10-15-2015

Chemometric and Bioinformatic Analyses of Cellular Biochemistry

Bradley Worley
University of Nebraska-Lincoln, bradley.worley@huskers.unl.edu

Chemometric and Bioinformatic Analyses of Cellular Biochemistry

by

Bradley Worley

A Dissertation

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Chemistry

Under the Supervision of Professor Robert Powers

Lincoln, Nebraska

October 2015

Chemometric and Bioinformatic Analyses of Cellular Biochemistry

Bradley Worley, Ph.D.

University of Nebraska, 2015

Advisor: Robert Powers

The amount of information collected and analyzed in biochemical and bioanalytical research has exploded over the last few decades, due in large part to the increasing availability of analytical instrumentation that yields information-rich spectra. Datasets from Nuclear Magnetic Resonance (NMR), Mass Spectrometry (MS), infrared (IR) or Raman spectroscopy may easily carry tens to hundreds of thousands of potentially correlated variables observed from only a few samples, making the application of classical statistical methods inappropriate, if not impossible. Drawing useful biochemical conclusions from these unique sources of data requires the use of specialized multivariate data handling techniques.

Unfortunately, proper implementation of many new multivariate algorithms requires domain knowledge in mathematics, statistics, digital signal processing, and software engineering *in addition to* analytical chemical and biochemical expertise. As a consequence, analysts using multivariate statistical methods were routinely required to chain together multiple commercial software packages and fashion small ad hoc software solutions to interpret a single dataset. This has been especially true in the field of NMR metabolomics, where no single software package, free or otherwise, was capable of completing all operations required to transform raw instrumental data into a set of validated, informative multivariate models. Therefore, while many powerful methods exist in published literature to statistically treat and model multivariate spectral data, few are readily available for immediate use by the community as a whole.

This dissertation describes the development of an end-to-end software solution for the handling and multivariate statistical modeling of spectroscopic data, called MVAPACK, and a set of novel spectral data acquisition, processing and treatment algorithms whose creation was expedited by MVAPACK. A final foray into the potential existence of $n - \pi^*$ interactions within proteins is also presented.

Acknowledgements

First and undeniably foremost, I extend my heartfelt gratitude to my family, whose encouragement and words of wisdom and guidance were arguably the most important contributor to my success in graduate school. Mom and Dad, you also sparked my scientific curiosity from a young age, and introduced me to so many scientific and intellectual opportunities. Elyssa, your encouragement and humor were always a much-needed reminder that light actually existed at the end of the tunnel I was passing through. I'm so proud of you, and I'm glad I have you as my sister.

Of course, no mention of family would be complete without my colleagues and mentors I had the privilege of working with on a daily basis. I'm happy the pressure cooker of graduate school brought us all closer than any ordinary coworkers. Dr. Martha Morton, your honest, uncensored insights on academic research, analytical chemical instrumentation and human interaction were always a pleasure to receive and will remain an invaluable aid in whatever future I choose to pursue. To Matt Shortridge, Jaime Stark, Jenni Copeland, Bo Zhang, Steve Halouska, Teklab Gebregiworgis, Darrell Marshall, Shulei Lei and Jonathan Catazaro, I'm thankful for the opportunities I had to share everything from scientific results to burdens of conscience with you. I would not have chosen to endure the rigours of graduate school if it were not for all of you. To the current Powers group, I challenge you to keep our truly remarkable comradery alive as the next round of junior students enters our clan. I have undoubtedly learned the most from all of you, and I hope those who join us next will share in that truly sustaining collaborative, friendly atmosphere that the Powers group thrives upon.

Finally, I wish to sincerely thank my advisor, Dr. Robert Powers, and the members of my supervisory committee, Drs. David Hage, Gerard Harbison, Eric Dodds, and Stephen Scott for providing me with an environment where I am free to pursue my intellectual curiosities, no matter how apparently unrelated they were at the time to this dissertation. Thank you all for your patient instruction and guidance during my time at the University of Nebraska.

List of Publications

This dissertation includes chapters that have been adapted from articles and communications published in peer-reviewed journals, as listed below:

Chapter 2

- B. Worley and R. Powers. Deterministic Multidimensional Nonuniform Gap Sampling. *Journal of Magnetic Resonance*, 2015

Chapter 3

- B. Worley and R. Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013
- B. Worley and R. Powers. A Sequential Algorithm for Multiblock Orthogonal Projections to Latent Structures. *Chemometrics and Intelligent Laboratory Systems*, 2015

Chapter 4

- B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014
- D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015
- B. Worley and R. Powers. PCA as a predictor of OPLS-DA model reliability. *Analytica Chimica Acta*, 2015

Chapter 5

- B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014

Chapter 6

- B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014

Chapter 7

- B. Worley, N. J. Sisco, and R. Powers. Statistical Removal of Background Signals from High-throughput ^{1}H NMR Line-broadening Ligand-affinity Screens. *Journal of Biomolecular NMR*, 63(4):53–58, 2015

Chapter 8

- B. Worley and R. Powers. Generalized Adaptive Intelligent Binning of Multiway Data. *Chemometrics and Intelligent Laboratory Systems*, 146:42–46, 2015

Chapter 9

- B. Worley and R. Powers. A Sequential Algorithm for Multiblock Orthogonal Projections to Latent Structures. *Chemometrics and Intelligent Laboratory Systems*, 2015

Chapter 10

- B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013

Chapter 11

- B. Worley, G. Richard, G. S. Harbison, and R. Powers. ^{13}C NMR Reveals No Evidence of $n - \pi^*$ Interactions in Proteins. *PLoS ONE*, 7(8):42075, 2012

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1
Introduction

> *As soon as the Analytical Engine exists, it will necessarily guide the future of science. Whenever any result is then sought by its aid, the question will then arise – by what course of calculation can these results be arrived at ... in the shortest time?*
>
> – Charles Babbage

1.1 Data Handling in Chemometrics

In analogy to biometrics, econometrics and psychometrics, the practice of chemometrics involves the extraction of chemically relevant information from measurements taken from chemical systems [33]. Naturally, this process of information extraction relies on the construction of mathematical models that describe a set of experimentally observed data, as well as statistical frameworks that assign degrees of belief (probabilities) to models, data, and their combinations:

$$\mathbf{D} = f(\mathbf{D}) + \mathbf{E}$$

In this highly generalized equation describing chemometric modeling, $\mathbf{D}$ is an experimentally measured dataset, $f(\mathbf{D})$ is a mathematical model that recapitulates $\mathbf{D}$, and $\mathbf{E}$ is the model "error", or information in the measured data that is not captured or described by the model. The ultimate goal of the analyst is to generate a set of measured data $\mathbf{D}$ and construct a model $f(\mathbf{D})$ that best describes that data (i.e. such that $||f(\mathbf{D})|| \gg ||\mathbf{E}||$). The above general equation describes a case of "unsupervised" chemometric modeling of the dataset $\mathbf{D}$, but analysts may also choose to construct a supervised model, where the data are used to predict a set of known responses $\mathbf{R}$:

$$\mathbf{R} = g(\mathbf{D} \mid \mathbf{R}) + \mathbf{E}'$$

where the model $g(\mathbf{D} \mid \mathbf{R})$ extracts information from the dataset that best describes $\mathbf{R}$, and the model error $\mathbf{E}'$ holds the differences between the known and modeled responses. Chemometrics is intimately connected with the chemical systems it aims to describe, and thus the exact choice of

mathematical model and statistical framework depends heavily on the particular problem, the data at hand, and the specific chemical information desired by the analyst.

As chemical systems under investigation increase in complexity, their chemometric description requires a proportionally increasing amount of measured data [33]. Biochemical systems at the levels of cellular metabolism and protein structure and function are arguably some of the most complex systems available for study by bioanalytical techniques, and demand vast amounts of spectral data in order to be suitably described by chemometric models [38, 18, 21, 5, 13, 3, 2, 24]. Proper handling of these large datasets requires novel tools and algorithms at each stage of the experimental process (Figure 1.1) in order to ensure maximal information extraction and minimal analyst errors.

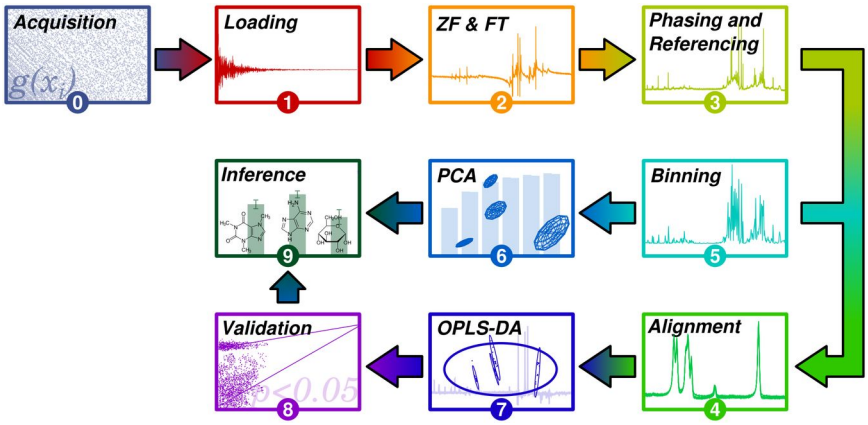

Figure 1.1: General Data Flow in Metabolomics.
Data in chemometric analyses of metabolism flows through this general graph, beginning at spectral data acquisition **(0)**, through to loading and processing of instrumental data **(1-3)**, further data treatment **(4-5)**, mathematical modeling **(6-7)** and model validation **(8)**, and terminating on extraction of chemical information **(9)**. In practice, this graph would be completely connected, and thus cyclic.

1.1.1 Acquisition

Nuclear Magnetic Resonance (NMR) spectroscopy is a popular analytical platform for chemometric analyses of protein structure and cellular metabolism, due to its ability to simultaneously report atomic-level details of the chemical environments and motional dynamics of ^{1}H, ^{13}C and ^{15}N nuclei in biomolecules [1, 20]. While the amount of information contained within one-dimensional (1D) NMR spectra is high, it is commonly held in a relatively narrow spectral width (e.g. -2.0 – 16 ppm for ^{1}H spectra). As a result, 1D ^{1}H NMR spectra of complex metabolite mixtures or biomacromolecules suffer from severe signal overlap that confounds analysis and interpretation.

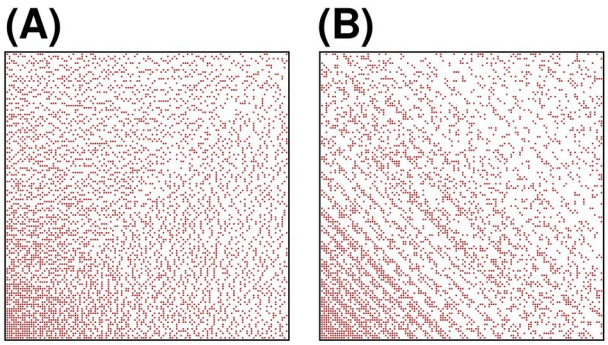

Figure 1.2: Example Nonuniform Sampling Schedules on a 2D Nyquist Grid.
Nonuniform sampling schedules produced by (**A**) stochastic and (**B**) deterministic subsampling of a two-dimensional Nyquist sampling grid. Comparisons of the performance of such schedules are made in Chapter 2.

Ever since the introduction of two-dimensional NMR methods by Jeener and Ernst [10, 25] and the popularization of three-dimensional methods for studying proteins by Bax and colleagues [23, 19], NMR spectroscopists have been leveraging ^{1}H–^{13}C and ^{1}H–^{15}N connectivities to spread biomolecular information from 1D ^{1}H spectra into two or more dimensions. While multidimensional experiments alleviate signal overlap, they require significantly more time to acquire than 1D spectra, as any D-dimensional experiment is effectively a $(D-1)$-dimensional array of 2^{D-1} one-dimensional experiments.[1] Time constraints imposed by throughput requirements, sample stability and instrumental maintenance have historically forced spectroscopists to harshly undersample their multidimensional datasets in the time domain, resulting in frequency domain digital resolutions much lower than the intrinsic line width of their samples [31, 28].

In order to move from this "sampling-limited" regime of data acquisition, the indirect dimensions of multidimensional NMR experiments may be nonuniformly sparsely sampled (Figure 1.2), reducing the time required for data collection while simultaneously enabling increased digital resolution [27]. When combined with non-Fourier reconstruction algorithms such as Maximum Entropy, ℓ_1-norm Minimization, and Multidimensional Decomposition [26], this technique of nonuniform sampling (NUS) is capable of producing high-quality, high-resolution multidimensional spectra in a fraction of the time required by traditional uniform sampling. However, the choice of which data points to subsample from a uniform Nyquist grid is nontrivial and has typically been made by random sampling methods [15, 22].

[1] The factor of 2^{D-1} arises from the fact that each dimension is collected in hypercomplex quadrature, as discussed in more detail in Chapter 3.

1.1.2 Processing and Treatment

More often than not, effective chemometric modeling of raw experimental data requires the data to be slightly modified from its original form. As an example, two commonly utilized soft bilinear modeling algorithms, principal component analysis (PCA, [17]) and partial least squares (PLS, [34]), analyze the eigenstructure of one or more data matrices, and require subtraction of the sample mean and scaling by the sample standard deviation in order to operate most effectively. When this modification is instrumentation-specific, it is referred to as *processing*; otherwise, it is considered a form of statistical *treatment*. The choice of which processing and treatment methods to apply to a given dataset $\mathbf{D}$ varies, depending on how the data were collected, which model $f(\mathbf{D})$ is used, and what information is sought from the model by the analyst.

Processing of NMR spectral datasets presents unique challenges to the analyst, as each spectrum is collected in hypercomplex quadrature [29] without absolute phase information. As a result, NMR spectra must be phase-corrected to maximize the real spectral component (cf. Chapter 3). When multiple spectra are processed as part of a statistical ensemble, any differences in phase *between* spectra become a contributing factor to undesirable within-group variation that inflates model errors. Thus, methods of phase-correcting multiple spectral observations are required when those observations will become inputs into multivariate modeling algorithms [36].

Once instrument-specific processing has been performed on a dataset $\mathbf{D}$, general statistical treatment operations may then be necessary, depending on the model function $f(\mathbf{D})$ being utilized. One commonly practiced method of preconditioning the eigenstructure of $\mathbf{D}$ for PCA and PLS, known as binning or bucketing, involves partitioning $\mathbf{D}$ into smaller signal-containing spectral regions and integrating or vectorizing those regions in order to achieve reduced data dimensionality. Multiple methods of binning one-dimensional datasets have been developed, ranging from naïve uniform subdivision algorithms [14, 30] to high-performance recursive methods [6, 8]. However, at the time of this writing, no methods of intelligently (non-uniformly) binning multidimensional datasets have been developed, essentially restricting bilinear PCA and PLS modeling to using 1D ^{1}H NMR spectral data in NMR chemometric studies of metabolism.

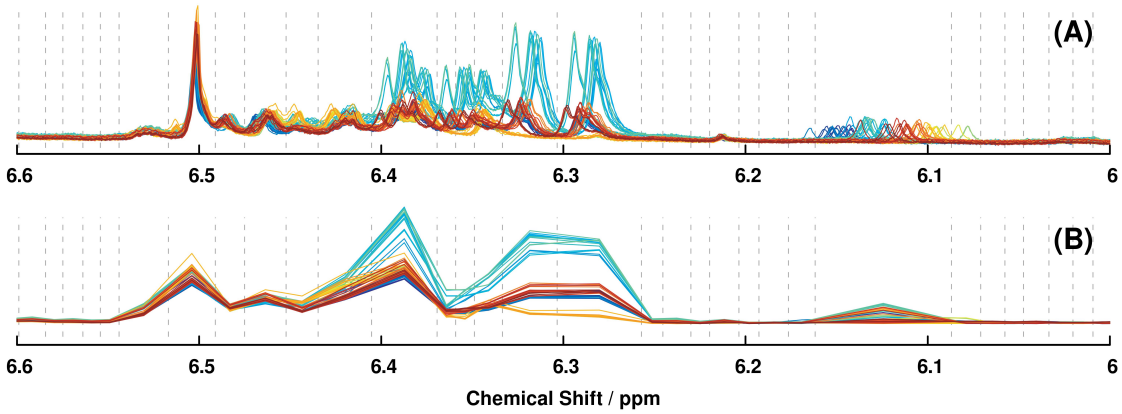

Figure 1.3: Example Binning Result from a 1D ^{1}H NMR Dataset.
Full-resolution (**A**) and adaptively intelligently binned (**B**) 1D ^{1}H NMR spectra from a chemometric study of brewed coffee roasts. Spectral color indicates the observation index, and dashed lines indicate bin boundaries. Further discussion of binning may be found in Chapters 3 and 6.

1.1.3 Modeling and Validation

Once a dataset $\mathbf{D}$ has been suitably processed and treated, a model $f(\mathbf{D})$ may be trained on its contents. Within chemometrics, principal component analysis (PCA) is undoubtedly the most routinely used modeling algorithm for describing relationships between multivariate spectral observations [4], because it provides an unbiased, simplified picture of the data in a low-dimensional "scores" space. The scores obtained from PCA models of spectral data are useful for determining statistical distances between experimental groups [7, 35], which are effective predictors of the reliability of any regression models that may be trained on the same data.

Another multivariate algorithm of equal popularity to PCA in chemometrics is partial least squares (PLS), which is used for solving regression and class discrimination problems on multivariate data [34]. While PLS provides a similar low-dimensional scores-space view of spectral observations, its true power in chemometrics lies in its ability to report "loadings", which are spectral contributions that predict a set of chemical properties.

The combination of PCA and PLS as a methodology for studying complex spectral datasets has proven highly useful in chemometrics, most notably so in the field of metabolomics [21]. However, analysts must take care when using models produced by these methods, as they have not been determined using standard (over-determined) least-squares methods and may over-fit a dataset at the expense of generality, which is required for broad inference [32]. Rigorous application of cross-

validation methods, including internal and external cross-validation [39, 11], response permutation testing [12] and CV-ANOVA [9], is required in order to ensure that trained multivariate models are reliable and generalizable to later measurements.

1.1.4 Inference

Once multivariate models have been trained and validated on a given dataset, they may finally be utilized for the extraction of chemical information from that dataset. Often, this process of inference revolves around the analysis of separations between one or more experimental groups in PCA or PLS scores space. Because scores-space separations are often used as justification for further costly experimentation, it is important to quantitatively measure these separations using proper statistical tools [35].

1.2 Summary of Work

By and large, this dissertation follows the logical flow of a data analyst in the field of NMR metabolomics, working from methods in compressed data acquisition, through a description of multivariate analysis techniques, to processing, treatment and validation of multivariate modeling results, and ending with a solution to a bioinformatics data handling problem: the correlation between high-resolution protein structure and backbone chemical shifts.

Chapter 2 begins by introducing a gap-based nonuniform sampling framework that provides several attractive advantages over traditional probability density-based nonuniform sampling methods. While most methods of generating nonuniform sampling schedules rely on randomly sampling from a specified weighting function that is defined over a Nyquist grid, this new method of gap sampling builds up schedules based on the value of a "gap equation" that specifies the spacing between sampled Nyquist grid points. The gap sampling framework is first defined, and comparisons in performance are made between specific forms of gap sampling and the stochastic Poisson-gap sampling method from Hyberts and Wagner [16].

A comprehensive description of the required data handling tasks – steps **(1-9)** in Figure 1.1 – in metabolic fingerprinting and untargeted metabolic profiling studies is provided within Chapter 3. Additional practical guidelines on the relationship between class separations in PCA scores space

and reliability of OPLS-DA models on the same data are also presented. Examples of applied multivariate analysis in metabolomics are given in Chapter 4.

Chapter 5 introduces the MVAPACK toolbox for chemometrics as a complete solution to the data handling problem in NMR- and MS-based metabolomics studies. Beginning with a set of raw free induction decays from an NMR spectrometer, analysts may now rapidly and easily generate validated multivariate models using rigorously tested and peer-reviewed routines in MVAPACK. As a result, both the turnaround time between data collection and interpretation and the likelihood of analyst error are dramatically reduced when using MVAPACK. The architecture and design rationale of the MVAPACK toolbox are discussed in this chapter.

Chapter 6 and Chapter 7 focus on a novel method of data processing (Phase-scatter Correction) and describe its application on datasets acquired from both metabolomics and high-throughput protein-ligand affinity screens. Chapter 8 introduces a novel method of data treatment (Generalized Adaptive Intelligent Binning) that enables the direct use of multidimensional data tensors in PCA and PLS modeling. Both phase-scatter correction and GAI-binning were developed within the MVAPACK toolbox, which was specifically designed for efficient management of NMR spectral data.

Chapter 9 introduces the multiblock orthogonal projections to latent structures (MB-OPLS) modeling method for handling predictive and non-predictive variation in a set of observed data matrices. Moving from modeling to inference, Chapter 10 describes a small set of portable utilities that generate statistically sound dendrograms of scores-space class relationships using both bootstrap-based and parametric methods.

Chapter 11 outlines the generation of a set of bioinformatic tools to analyze the relationship between the geometry of interacting pairs of carbonyls in protein backbones and their ^{13}C chemical shift values [37]. These tools, combined with quantum chemical computations, provide strong evidence for the nonexistence of $n-\pi^*$ interactions between these carbonyl groups in native protein structures.

Finally, Chapter 12 summarizes the solutions provided herein to a set of chemometrics and bioinformatics data handling problems and discusses challenges and avenues of effort that will be required to solve future problems of the same kind.

1.3 References

[1] A. Abragam. *The Principles of Nuclear Magnetism*. Oxford University Press, 1961.

[2] J. M. Baker, J. L. Ward, and M. H. Beale. Combined NMR and flow injection ESI-MS for Brassicaceae metabolomics. *Methods in Molecular Biology*, 860:177–191, 2012.

[3] G. A. Barding, D. J. Orr, T. Fukao, J. Bailey-Serres, and C. K. Larive. Multi-platform metabolomics: Combining NMR and GC-MS to give a deeper understanding of the rice metabolome. *Journal of the American Chemical Society*, 243, 2012.

[4] R. Bro and A. K. Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

[5] H. Chen, Z. Pan, N. Talaty, D. Raftery, and R. G. Cooks. Combining desorption electrospray ionization mass spectrometry and nuclear magnetic resonance for differential metabolomics without sample preparation. *Rapid Communications in Mass Spectrometry*, 20(10):1577–1584, 2006.

[6] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson. Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, 85(1):144–154, 2007.

[7] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[8] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[9] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[10] R. Ernst. Multidimensional NMR. *Chimia*, 29:179–187, 1975.

[11] P. Eshghi. Dimensionality choice in principal components analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems*, 130:6–13, 2014.

[12] P. Golland, F. Liang, S. Mukherjee, and D. Panchenko. Permutation tests for classification. *Learning Theory*, 1:501–515, 2005.

[13] J. Han, R. M. Danell, J. R. Patel, D. R. Gumerov, C. O. Scarlett, J. P. Speir, C. E. Parker, I. Rusyn, S. Zeisel, and C. H. Borchers. Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. *Metabolomics*, 4(2):128–140, 2008.

[14] M. Hedenstrom, S. Wiklund, B. Sundberg, and U. Edlund. Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometrics and Intelligent Laboratory Systems*, 92(2):110–117, 2008.

[15] J. C. Hoch, M. W. Maciejewski, and B. Filipovic. Randomization improves sparse sampling in multidimensional NMR. *Journal of Magnetic Resonance*, 193(2):317–320, 2008.

[16] S. G. Hyberts, K. Takeuchi, and G. Wagner. Poisson-Gap Sampling and Forward Maximum Entropy Reconstruction for Enhancing the Resolution and Sensitivity of Protein NMR Data. *Journal of the American Chemical Society*, 132(7):2145–2147, 2010.

[17] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[18] L. E. Kay. NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance*, 173(2):193–207, 2005.

[19] L. E. Kay, D. A. Torchia, and A. Bax. Backbone Dynamics of Proteins as Studied by ^{15}N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease. *Biochemistry*, 28:8972–8979, 1989.

[20] M. H. Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, 2008.

[21] J. C. Lindon, J. K. Nicholson, E. Holmes, and J. R. Everett. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance*, 12(5):289–320, 2000.

[22] M. W. Maciejewski, H. Z. Qui, I. Rujan, M. Mobli, and J. C. Hoch. Nonuniform sampling and spectral aliasing. *Journal of Magnetic Resonance*, 199(1):88–93, 2009.

[23] D. Marion, L. E. Kay, S. W. Sparks, D. A. Torchia, and A. Bax. Three-Dimensional Heteronuclear NMR of ^{15}N-Labeled Proteins. *Journal of the American Chemical Society*, 111:1515–1517, 1989.

[24] D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015.

[25] A. Maudsley and R. Ernst. Indirect detection of magnetic resonance by heteronuclear two-dimensional spectroscopy. *Chemical Physics Letters*, 50(3):368–372, 1977.

[26] M. Mobli and J. C. Hoch. Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 83C:21–41, 2014.

[27] D. Rovnyak, D. P. Frueh, M. Sastry, Z. Y. J. Sun, A. S. Stern, J. C. Hoch, and G. Wagner. Accelerated acquisition of high resolution triple-resonance spectra using non-uniform sampling and maximum entropy reconstruction. *Journal of Magnetic Resonance*, 170(1):15–21, 2004.

[28] D. Rovnyak, J. C. Hoch, A. S. Stern, and G. Wagner. Resolution and sensitivity of high field nuclear magnetic resonance spectroscopy. *Journal of Biomolecular NMR*, 30(1):1–10, 2004.

[29] A. D. Schuyler, M. W. Maciejewski, A. S. Stern, and J. C. Hoch. Formalism for hypercomplex multidimensional NMR employing partial-component subsampling. *Journal of Magnetic Resonance*, 227:20–24, 2013.

[30] S. A. A. Sousa, A. Magalhaes, and M. M. C. Ferreira. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122:93–102, 2013.

[31] T. Szyperski, D. C. Yeh, D. K. Sukumaran, H. N. B. Moseley, and G. T. Montelione. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. *PNAS*, 99(12):8009–8014, 2002.

[32] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[33] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[34] S. Wold, E. Johansson, and M. Cocchi. *PLS: Partial Least Squares Projections to Latent Structures*. KLUWER ESCOM Science Publisher, 1993.

[35] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013.

[36] B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014.

[37] B. Worley, G. Richard, G. S. Harbison, and R. Powers. ^{13}C NMR Reveals No Evidence of $n-\pi^*$ Interactions in Proteins. *PLoS ONE*, 7(8):42075, 2012.

[38] K. Wütrich, G. Wider, G. Wagner, and W. Braun. Sequential resonance assignments as a basis for determination of spatial protein structures by high resoltuion proton nuclear magnetic resonance. *Journal of Molecular Biology*, 155(3):311–319, 1982.

[39] Q. S. Xu and Y. Z. Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

Chapter 2
Multidimensional Nonuniform Gap Sampling

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.

– John von Neumann

2.1 Introduction

The use of nonuniform sampling in multidimensional NMR is rapidly becoming standard practice in most biomolecular solution-state experiments, thanks in large part to recent developments in fast reconstruction algorithms, novel sampling schemes, and the continually declining cost of computing power [18]. The potential benefits of collecting a subset of the full Nyquist grid – including increased sensitivity and signal-to-noise, improved resolution, and reduced experiment time – have received significant attention [21, 22, 13, 10, 20] in recent years as a consequence.

One intriguing result of recent investigations into the parameters of NUS experiments is the use of random deviates for generating sampling schedules [5]. In fully random sampling schemes, a subset of Nyquist grid points is drawn from a probability density function that varies over the grid, producing a sampling schedule with a desired distribution of points. Common fully random sampling schemes utilize uniform, exponential, Gaussian and envelope-matched probability densities [20, 23]. While randomization is a simple means of reducing the artifacts due to aliasing of nonuniformly spaced samples, it turns the already complex task of schedule generation into that of selecting a schedule from an ensemble of possibilities, each of which performs differently in practice [11, 17]. Several ad hoc metrics have been proposed to assess the relative performance sampling schedules, but no universally accepted metric exists to guide the selection of a stochastic schedule from its ensemble [17, 1]. Without a priori knowledge of the frequency and decay rate distributions of the signals to be measured, it is difficult to reliably quantify sampling schedule performance [18, 23]. As a result, numerous recent attempts have been made to reduce or remove pseudorandom seed-dependent variability from nonuniform sampling algorithms [14, 11, 4, 17]. Such efforts are an important step to-

wards increasing the practical utility of nonuniform sampling in everyday spectroscopic experiments.

One prominent method designed to reduce seed-dependent variability in pseudorandomly constructed schedules in Poisson-gap sampling. Through an empirical analysis of Forward Maximum Entropy (FM) reconstructions of randomly sampled data, Hyberts et al. proposed the use of constrained Poisson random deviates to define the *gaps* between sampled points in a Nyquist grid [11]. The FM reconstruction residuals of these so-named Poisson-gap schedules exhibited a markedly lower dependence on seed value than unconstrained random sampling methods. While Poisson-gap sampling yields high-quality reconstructions of NUS spectral data, its average behavior is not well-understood, its implementation for multidimensional Nyquist grids is unclear [8, 9, 7], and its relationship – if any – to fully random sampling is unknown. To meet this need, this work describes in detail the deterministic generation of sinusoidally weighted multidimensional gap schedules that model the average behavior of stochastic Poisson-gap (PG) sampling. An expectation sampling probability distribution is also derived that reflects the average weighting obtained using one-dimensional Poisson-gap sampling schedules.

Among the myriad of different sampling schemes proposed for NUS data collection [15], burst-mode sampling similarly concerns itself with gaps between sampled grid points. Unlike Poisson-gap sampling, which aims to minimize the *length* of gaps, burst-mode sampling aims to minimize the *number* of gaps while keeping the effective dwell time low [16]. We leverage the complementarity of burst-mode and Poisson-gap sampling in our deterministic gap sampling algorithm to describe a novel sampling scheme that simultaneously seeks to bias sample collection to early times, minimize the number of long gaps between densely sampled regions, and minimize the largest gap length in the schedule. The resulting method, called sine-burst (SB) sampling, exhibits the high performance of Poisson-gap sampling while retaining the bijective mapping between inputs and outputs offered by deterministic methods.

2.2 Theory

2.2.1 Poisson-gap Sequences

Gap schedules on a one-dimensional Nyquist grid are effectively finite integer sequences, computed from the following recurrence relation:

$$x_{i+1} = x_i + \lfloor g(x_i) \rfloor + 1 \tag{2.1}$$

where x_i is the grid index of the i-th term in the sequence and $g(x_i)$ is the "gap equation" that defines the distance between terms. The first term in the sequence, x_1, is set to 1 and subsequent terms are computed until their value exceeds N, the size of the grid. The gap equation $g(x)$ may be any non-negative function, and may be loosely interpreted as inversely related to the local sampling density at the grid index x_i. Thus, when the gap equation equals zero for all grid indices, gap sampling will yield a uniformly sampled grid.

Poisson-gap sequences treat the gap equation as a Poisson random deviate having a rate parameter that varies as either a quarter- or half-sinusoid over the grid indices:

$$g_{PG}(x_i) \sim \text{Pois}\left\{\Lambda \sin\left(\frac{\pi}{2}\theta_i\right)\right\} \tag{2.2}$$

where Λ is a scaling factor that determines the global sampling density and θ_i is the fractional grid index that varies from 0 to 1 over the grid extents:

$$\theta_i = \frac{x_i}{N} \tag{2.3}$$

In all following descriptions of Poisson-gap methods, we shall restrict our attention to rate parameters which vary as quarter-sinusoids, where the fractional grid index is multiplied by a factor of one-half π. This choice of sinusoidal weight produces schedules that are heavily biased to earlier grid points. Using a factor of π produces half-sinusoidal rate parameters and schedules that are more densely sampled at both early and late grid points.

Because the expected value of a Poisson distribution is equal to its rate parameter, one may trivially construct a deterministic sinusoidally weighted gap sampler (sine-gap; SG) by setting the gap

equation equal to the scaled quarter-sinusoid from equation 2.2, as follows:

$$g_{SG} = \Lambda \sin\left(\frac{\pi}{2}\theta_i\right) \tag{2.4}$$

By construction, gap sampling schedules computed according to g_{SG} will describe the average behavior of g_{PG}. This is readily verified in one dimension by generating a sufficiently large set of stochastic Poisson-gap schedules and comparing the mean value of each sequence term to that of a sine-gap schedule (Figure 2.1). Sine-gap schedules lie centrally within the Poisson-gap ensemble, while other schedules unrelated to Poisson-gap deviate substantially from the confidence region of the ensemble.

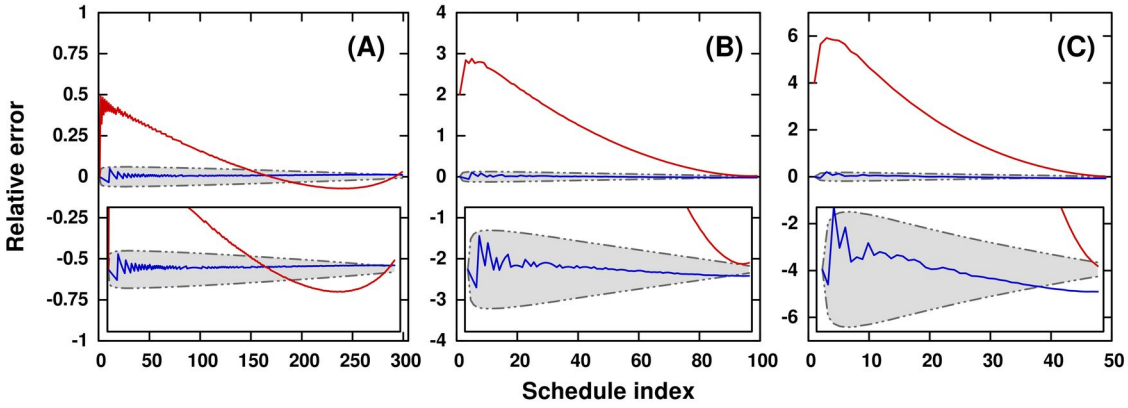

Figure 2.1: Agreement between Poisson-gap and sine-gap Sequences.
Relative errors between sine-gap (blue lines) and deterministic exponential (red lines) schedules with respect to the average Poisson-gap schedule at (**A**) 30% density, (**B**) 10% density and (**C**) 5% density. Confidence intervals indicating plus or minus one standard deviation of the Poisson-gap ensemble are shown as gray shaded regions. The vertical axes of all inset plots range from -0.2 to 0.2. Because the sine-gap schedules describe the average behavior of the Poisson-gap equation, they lie centrally within the Poisson-gap ensemble, while any other schedule unrelated to Poisson-gap (e.g. exponential) does not.

2.2.2 Multidimensional Gap Sampling

Gap schedules on a Nyquist grid having at least two dimensions are generated by placing multiple one-dimensional sub-schedules onto the grid, each with a different direction and offset from the grid origin. In practice, this process is accomplished recursively, with planes built up from vectors, cubes built up from planes, and so forth. Initially, recursion begins on the entire grid. At each level of recursion, sub-grids are constructed by "masking off" each available grid direction in turn and constructing the remaining unmasked directions. For example, a three-dimensional xyz cube will be constructed from repeated sequences of yz, xz and xy planes, and each xy plane will be

constructed from repeated sequences of y and x vectors. Once a round of sub-grid construction has been performed along each direction, the sub-grid offset is incremented and the process is repeated until no more sub-grids remain at the current level of recursion. The following executable pseudocode provides a more precise definition of the recursive gap sampling algorithm:

Algorithm 2.1 Multidimensional Gap Sampling Algorithm

```
def build(N, origin, mask):
  D = len(N)
  if sum(mask) == 1:
    direction = mask.index(1)
    dirstring = ['x', 'y', 'z'][direction]
    print('sequence along ' + dirstring + ' at origin ' + origin)
    return

  suborigin = [0,] * D
  submask = [0,] * D
  done = False
  offset = 0

  while not done:
    done = True
    for direction in range(D):
      if mask[direction] != 1 or offset >= N[direction]:
        continue

      done = False
      for d in range(D):
        if d != direction and mask[d] == 1:
          submask[d] = 1
        else:
          submask[d] = 0

        if d == direction:
          suborigin[d] = offset
        else:
          suborigin[d] = origin[d]

      build(N, suborigin, submask)
    offset = offset + 1

build([8, 4, 4], [0, 0, 0], [1, 1, 1])
```

Creation of multidimensional gap schedules requires a slight modification to the fractional index, which now assumes the following form:

$$\theta_i = \frac{x_i + \sum_{d=1}^{D} O_d}{\sum_{d=1}^{D} N_d} \tag{2.5}$$

where O_d and N_d are the origin and grid size along direction d, respectively. The above equation is referred to as "ADD" mode in the context of stochastic Poisson-gap sampling, and effectively results in multidimensional schedules that exhibit triangular forms [7]. It is worthy of mention that, in the one-dimensional case, equation 2.5 reduces to equation 2.3.

Finally, whether the Nyquist grid is one- or many-dimensional, a value of the global scaling factor Λ must be determined that yields the desired number of sampled grid points. Our gap sampling implementation, like the existing Poisson-gap method, iteratively rebuilds new schedules until Λ has been suitably optimized. The published implementation uses a heuristic search method that adjusts Λ based on the relative difference between the desired and obtained global sampling density at each iteration.

2.2.3 Burst Augmentation

Recent statistical descriptions of the discrete Fourier transform have shown that the bandwidth of a nonuniformly sampled signal is related to the greatest common factor of the gaps between sampled grid points [2]. One proposed method of increasing bandwidth and reducing artifacts in NUS data is to sample in multiple short bursts having zero gap length [16]. Using gap sampling, this may be accomplished by modulating the gap equation between zero and its maximum value several times over the Nyquist grid, like so:

$$g_{SB}(x_i; d) = \Lambda \sin\left(\frac{\pi}{2}\theta_i\right) \sin^2\left(\frac{\pi}{4}N_d\theta_i\right) \tag{2.6}$$

The sine-burst gap equation g_{SB} combines the sinusoidal forward-biasing and minimized gap lengths of Poisson-gap sampling with the minimized effective dwell time of burst-mode sampling, and does not require the use of random deviates to achieve reasonable artifact suppression.

2.2.4 Expectation Sampling Distributions

One disadvantage of stochastic gap equations is that they provide no direct measure of how likely each Nyquist grid point is to be sampled. While one may speculate on the approximate weighting obtained by a given gap equation, quantitation of the expectation of the sampling distribution requires the construction and averaging of a large number of sampling schedules (cf. Figures 2.2, 2.3 and 2.4). Fortunately, the expectation sampling distribution of a given gap equation may be analytically

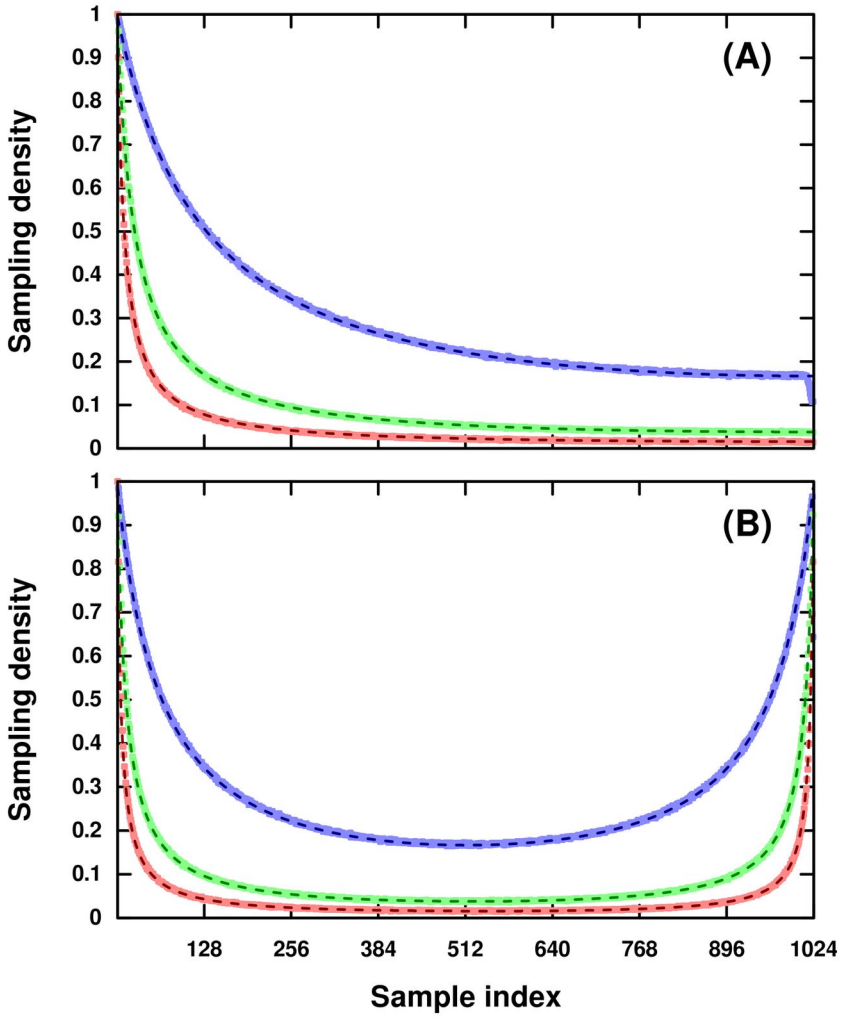

Figure 2.2: Poisson-gap 1D Expectation Sampling Distributions.
Expectation sampling distributions computed by averaging 50,000 one-dimensional Poisson-gap schedules of varying densities, with quarter-sinusoidal (**A**) and half-sinusoidal (**B**) weightings. The lighter blue, green and red points were computed from schedules having 30%, 10% and 5% sampling density, respectively. The dashed lines overlaid on each set of points correspond to the analytic sampling distribution (equation 2.11) derived from g_{PG}. Values of Λ were 5.0, 25.4 and 62.9 for 30%, 10% and 5% sampling density, respectively.

obtained by computing the probability of sampling each point on the grid using a recursive formula. We define an expectation sampling distribution $p(i)$ that varies over a one-dimensional Nyquist grid of N points as follows:

$$p(i) = \sum_{k=1}^{i-1} p(i \mid i-k)p(i-k) \tag{2.7}$$

where $p(i \mid i-k)$ is the probability of grid point i being emitted from grid point $i-k$, which requires a gap size of $k-1$:

$$p(i \mid i-k) = \Pr\{\lfloor g(i-k) \rfloor = k-1\} \tag{2.8}$$

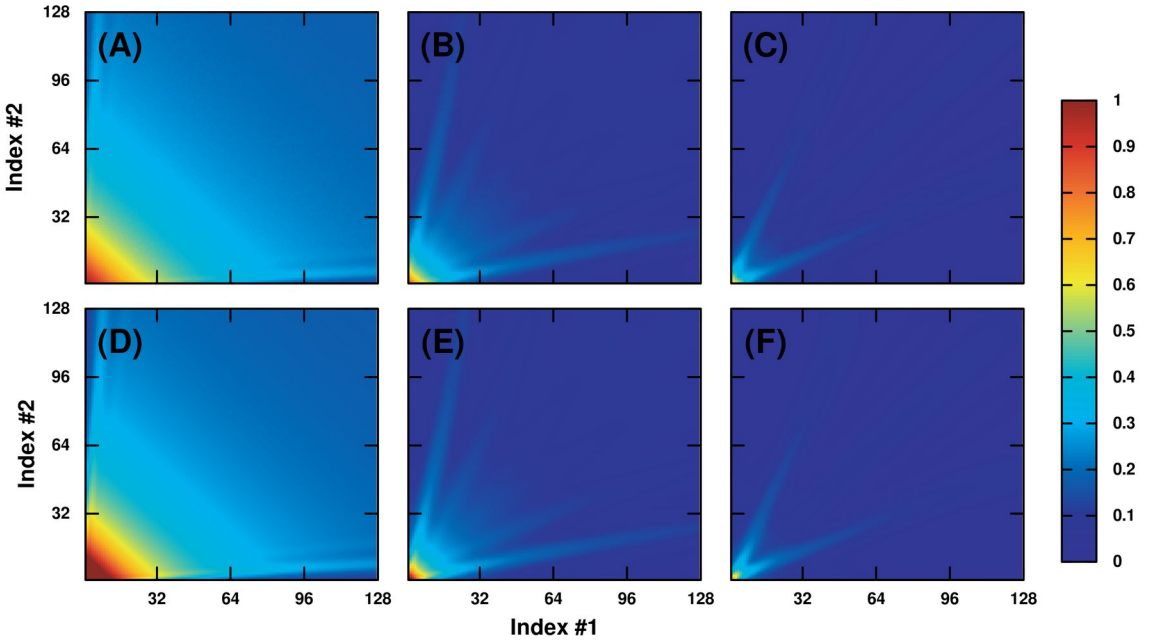

Figure 2.3: Poisson-gap 2D Expectation Sampling Distributions.
Expectation sampling distributions of two-dimensional Poisson-gap schedules computed in strict accordance to Algorithm 2.1. Top panels were produced by averaging 50,000 two-dimensional schedules, and bottom panels were computed using equation 2.12 with appropriate substitutions of the Poisson probability mass function. Sampling densities of 30%, 10% and 5% are shown in panels (**A**, **D**), (**B**, **E**) and (**C**, **F**), respectively.

In other words, the probability of sampling any given grid point is the weighted sum of the probabilities of arriving at that point from all prior points. In the case of Poisson-gap sampling, the gap equation is a Poisson random deviate:

$$\Pr\{\lfloor g(x_i) \rfloor = k-1\} = \frac{\lambda(x_i)^{k-1}}{(k-1)!} e^{-\lambda(x_i)} \tag{2.9}$$

where the rate parameter $\lambda(x_i)$ varies sinusoidally over the Nyquist grid:

$$\lambda(m) = \Lambda \sin\left(\frac{\pi m}{2N}\right) \tag{2.10}$$

By combining the above four equations, we arrive at the sampling distribution of a one-dimensional

Poisson-gap sequence:

$$p(i) = \sum_{k=1}^{i-1} \frac{\Lambda^{k-1}}{(k-1)!} \sin^{k-1}\left(\frac{\pi[i-k]}{2N}\right) \exp\left\{-\Lambda \sin\left(\frac{\pi[i-k]}{2N}\right)\right\} p(i-k) \tag{2.11}$$

As in the case of gap sampling, the sampling distribution produced by equation 2.11 is parameterized only by the scaling factor Λ, where larger values yield more forward-biased schedules (Figure 2.2). We refer to this equation as the "expectation" Poisson-gap sampling distribution because it describes the expected value of the probability of sampling any Nyquist grid point, and is not itself useful for generating schedules that obey g_{PG}.

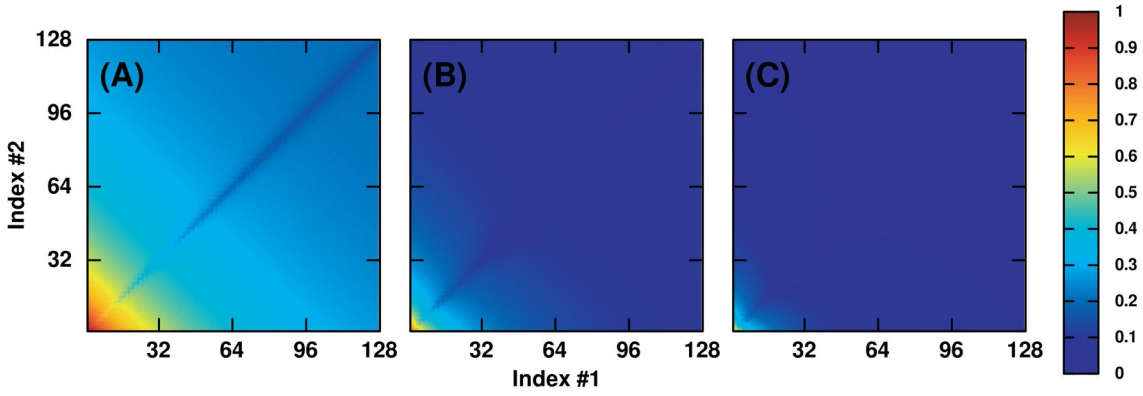

Figure 2.4: Poisson-gap 2D Expectation Sampling Distributions.
Expectation sampling distributions computed by averaging 50,000 two-dimensional Poisson-gap schedules generated using code provided by Hyberts and Wagner. Sampling densities of 30%, 10% and 5% are shown in panels (**A**), (**B**) and (**C**), respectively.

2.2.5 Multidimensional Expectation Sampling Distributions

Extension of equation 2.11 to compute the expectation sampling distributions of stochastic gap equations in two or more dimensions follows from the fact that sampling along each direction is essentially independent of other directions within the presented gap sampling framework. As a consequence, the probability of sampling any multidimensional grid point is therefore the sum of sampling that point along each grid direction. For a two-dimensional grid (i_1, i_2), the expectation

sampling distribution is the sum of the probability matrices $p_1(i_1, i_2)$ and $p_2(i_1, i_2)$:

$$p_1(i_1, i_2) = \sum_{k=1}^{i_1-1} p(i_1, i_2 \mid i_1 - k, i_2) p_1(i_1 - k, i_2)$$
$$p_2(i_1, i_2) = \sum_{k=1}^{i_2-1} p(i_1, i_2 \mid i_1, i_2 - k) p_2(i_1, i_2 - k)$$
$$p(i_1, i_2) = p_1(i_1, i_2) + p_2(i_1, i_2) \tag{2.12}$$

Figure 2.3 illustrates the expectation Poisson-gap sampling distribution on two-dimensional Nyquist grids. It is important to note that the Poisson-gap sampler originally proposed by Hyberts et al. does not strictly follow Algorithm 2.1, because its sampling of each dimension is dependent upon which points in other dimensions have been previously sampled. This divergence between multidimensional Poisson-gap and Poisson-gap constructed using Algorithm 2.1 is observed by comparison of Figures 2.3 and 2.4, and is only truly apparent at very low sampling densities.

2.3 Materials and Methods

2.3.1 Generation of Deterministic Schedules

Deterministic sine-gap and sine-burst schedules were constructed using a small C program that implements the presented gap sampling algorithm described above. Schedules were generated at 30%, 10% and 5% sampling densities on one-dimensional grids having 1,024 points and two-dimensional grids having 64×64 and 128×128 points. The first and third rows of Figure 2.5 show the deterministic schedules resulting from g_{SG} and g_{SB} at 30% density on 128×128 grids, respectively.

2.3.2 Generation of Stochastic Schedules

Poisson-gap schedules were constructed using Java source code authored and provided by Hyberts et al. for generating multidimensional schedules (`http://gwagner.med.harvard.edu/intranet/hmsIST/gensched_old.html`). A small command-line wrapper was written to provide direct access to the core schedule generation functions without use of the graphical interface. Fifty thousand schedules were computed at each of the sampling densities and grid sizes listed in Subsection 2.3.1. Each schedule was generated with a unique, large, odd-valued seed number to ensure the broadest possible sampling of the PG ensemble. The second row of Figure 2.5 shows a representative two-dimensional Poisson-gap schedule at 30% sampling density.

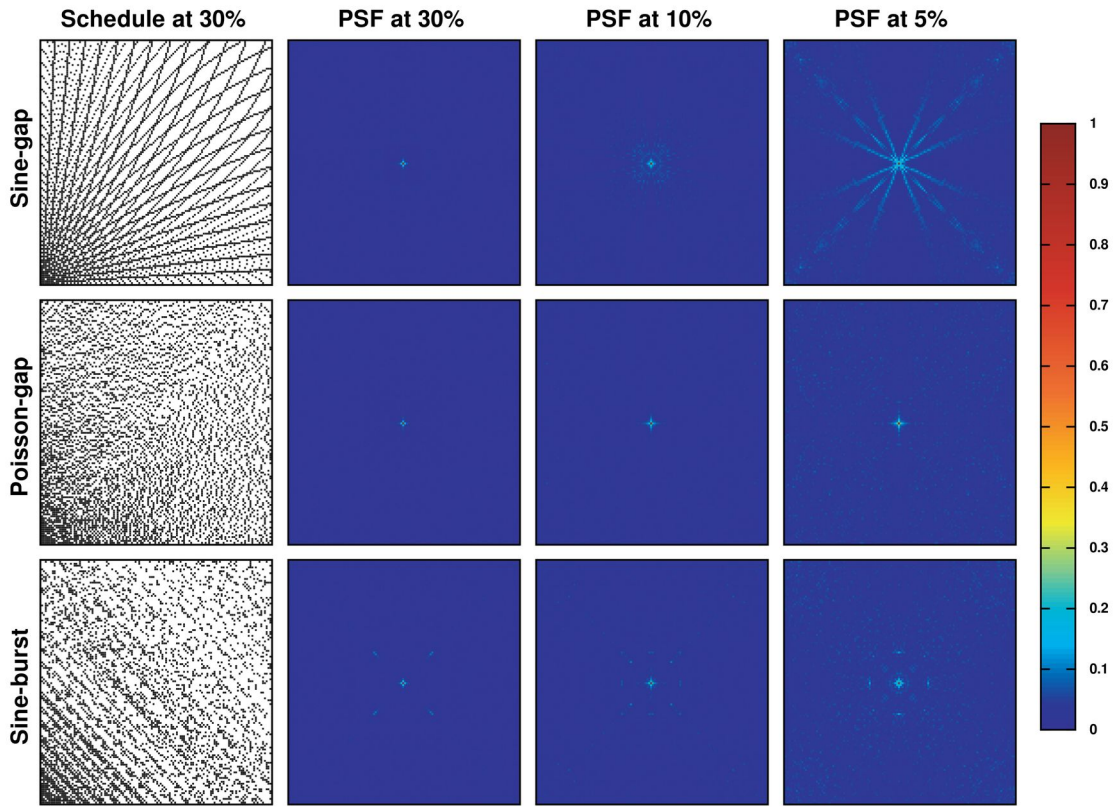

Figure 2.5: Comparison of Gap Sampling Schedules.
Comparison of sine-gap, Poisson-gap and sine-burst sampling schedules and their resulting point-spread functions at varying sampling densities, indicating close agreement between all methods. The increased artifact intensity in the sine-gap schedule at 5% sampling density is likely due to slightly increased regularity of sampled grid points, which is reduced by Poisson-gap and sine-burst sampling. Grid sizes and point spread function colorings are the log-scaled versions of those found in Figure 1 of [6] in order to emphasize low-intensity sampling artifacts.

2.3.3 Spectral Data Collection

A high-resolution 2D ^{1}H–^{15}N HSQC NMR spectrum was collected at a temperature of 298.0 K on a sample of uniformly [^{15}N, ^{13}C]-labeled ubiquitin in aqueous phosphate buffer at pH 6.5. Data were acquired on a Bruker Avance III HD 700 MHz spectrometer equipped with a 5 mm inverse quadruple-resonance (^{1}H, ^{13}C, ^{15}N, ^{31}P) cryoprobe with cooled ^{1}H and ^{13}C channels and a z-axis gradient. A 2D gradient-enhanced ^{1}H–^{15}N HSQC spectrum with improved sensitivity [12, 19] was collected with 16 scans and 32 dummy scans over a uniform grid of 2,048 and 1,024 hypercomplex points along the ^{1}H and ^{15}N dimensions, respectively. Spectral widths were set to 3,293±4,209 Hz along ^{1}H and 8,514±1,419 Hz along ^{15}N. The spectrum was windowed with a squared-cosine function, Fourier-transformed and phase-corrected along ^{1}H to produce a half-transformed spectrum for

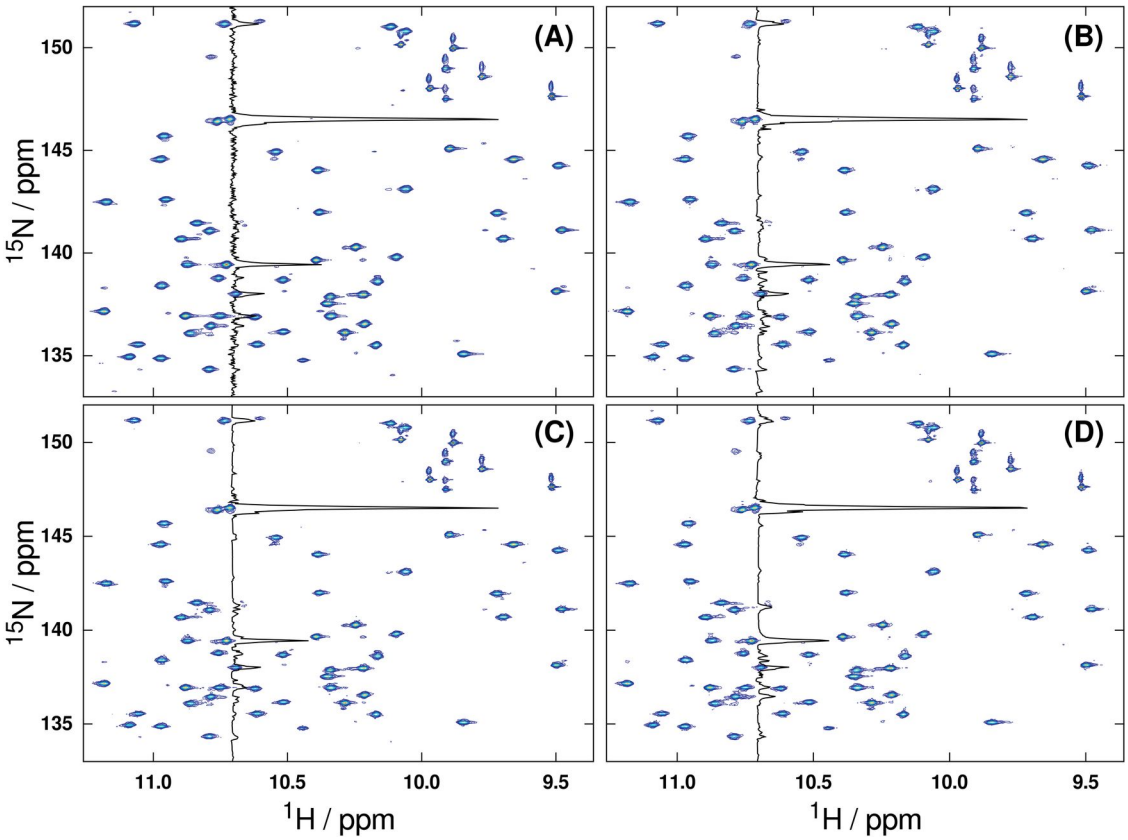

Figure 2.6: Comparison of HSQC Spectral Reconstructions.
Uniformly sampled (**A**) and IST reconstructed (**B**–**D**) 2D ^{1}H–^{15}N HSQC spectra of ubiquitin, indicating nearly equivalent performance of all three gap sampling methods at low (5%) sampling density. Spectra shown in (**B**) through (**D**) were reconstructed from nonuniformly subsampled copies of (**A**) using Poisson-gap (**B**), sine-gap (**C**), and sine-burst (**D**) schedules, respectively. All spectra are plotted with identical contour levels.

IST reconstruction analysis (vide infra), and subsequently windowed and Fourier-transformed along ^{15}N to yield the "true" uniformly sampled 2D ^{1}H–^{15}N HSQC spectrum. Figure 2.6 compares the true HSQC spectrum with representative IST reconstructions after subsampling by Poisson-gap, sine-gap and sine-burst schedules.

In addition, a 3D HNCA NMR spectrum was collected on the same uniformly [^{15}N, ^{13}C]-labeled ubiquitin sample. The spectrum was collected at 298.0 K with 16 scans and 32 dummy scans over a uniform grid of 1,024×64×64 hypercomplex points along the ^{1}H, ^{15}N and ^{13}C dimensions, respectively. Spectral windows were set to 3,293±4,209 Hz along ^{1}H, 8,514±1,419 Hz along ^{15}N, and 9,508±2,818 Hz along ^{13}C. The spectrum was windowed with a squared-cosine, Fourier-transformed and phase-corrected along ^{1}H to produce an F_3-transformed spectrum for IST reconstruction anal-

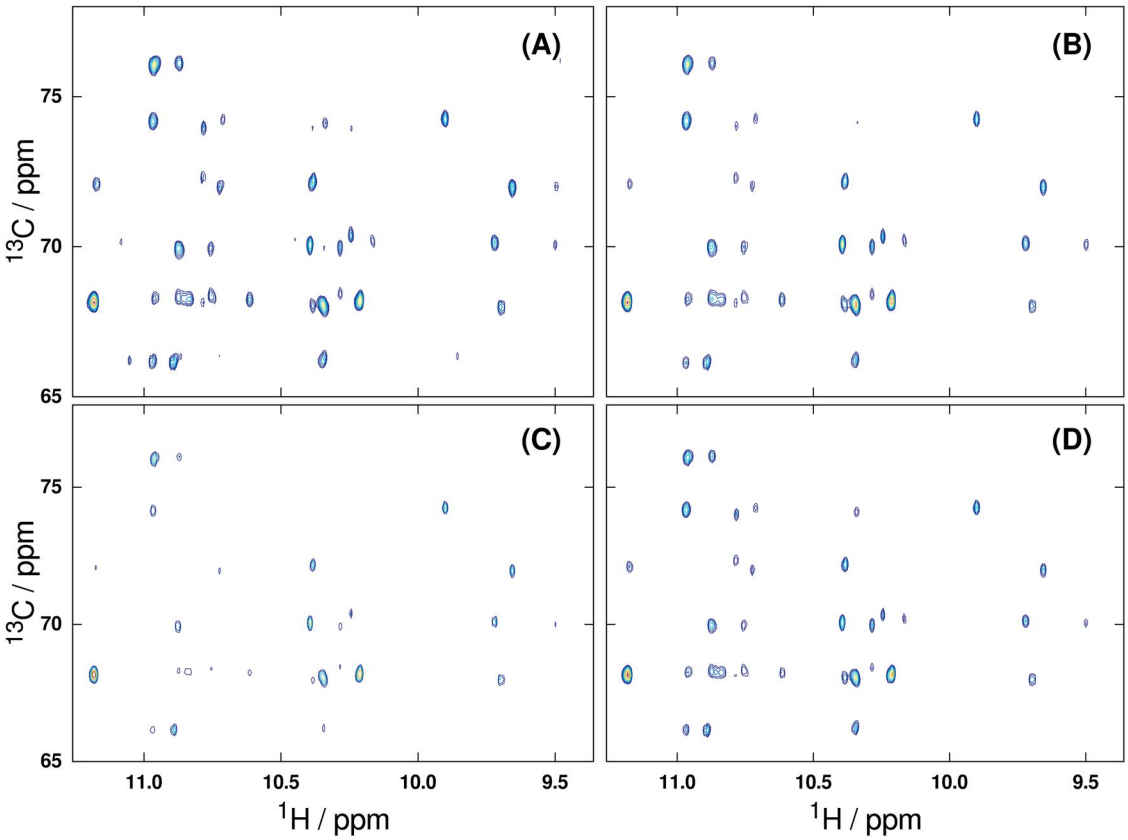

Figure 2.7: Comparison of HNCA Spectral Reconstructions.
Uniformly sampled (**A**) and IST reconstructed (**B**–**D**) 3D HNCA spectra of ubiquitin at low (5%) sampling density, projected along the ^{15}N dimension. Spectra shown in (**B**) through (**D**) were reconstructed from nonuniformly subsampled copies of (**A**) using Poisson-gap (**B**), sine-gap (**C**), and sine-burst (**D**) schedules, respectively. While sine-gap sampling (**C**) fails to adequately reproduce the spectrum due to its high sampling coherence, sine-burst sampling yields an essentially identical result to Poisson-gap sampling. All spectra are plotted with identical contour levels.

ysis, and subsequently windowed and Fourier transformed along ^{15}N and ^{13}C to yield the "true" uniformly sampled 3D HNCA spectrum. Figure 2.7 compares ^{1}H–^{13}C projections of the true HNCA with those of representative IST reconstructions after subsampling by Poisson-gap, sine-gap and sine-burst schedules.

2.3.4 Computation of Performance Metrics

All computational analyses were performed using in-house developed C programs. An implementation of the hypercomplex algebra described by Schuyler et al. [24] was used to perform all spectral data processing. Iterative soft thresholding (IST) reconstructions of subsampled spectra were performed using the algorithm described by Stern et al. [25, 26]. Impulse sets were generated for each

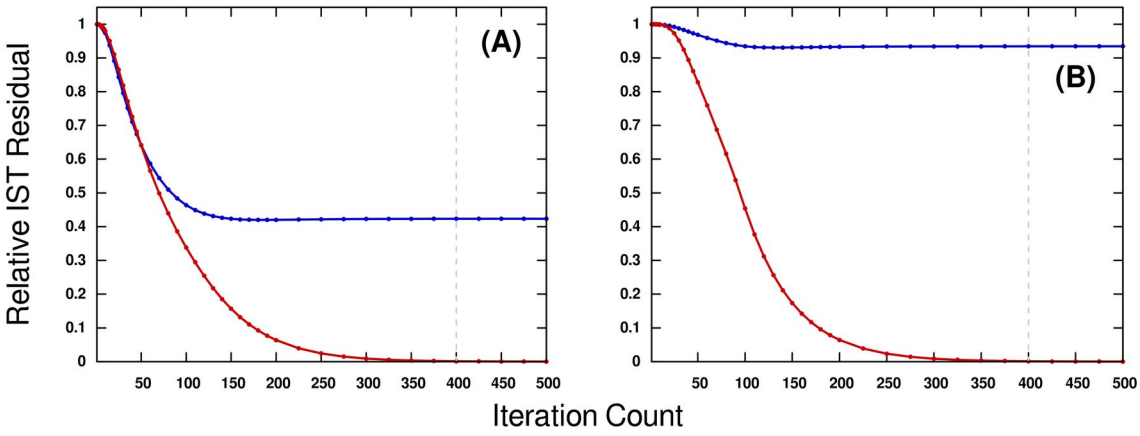

Figure 2.8: Convergence of IST Reconstruction Residuals.
Convergence analysis of IST reconstructions of (**A**) ^{1}H–^{15}N HSQC F_1 traces and (**B**) HNCA $F_2 - F_1$ planes. Both spectra were nonuniformly subsampled with a 5% density sine-gap schedule prior to IST. Relative ℓ_2 reconstruction errors computed from all data points are shown by blue lines, and errors computed only from initially sampled data points are shown by red lines. Grey dashed lines indicate the iteration count at which all IST reconstructions were performed to generate histograms of ℓ_2 errors.

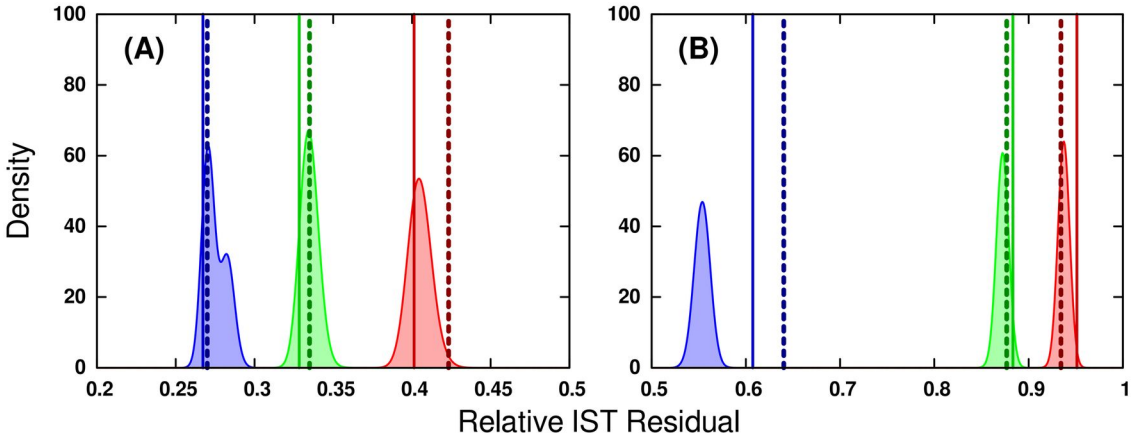

Figure 2.9: Relative IST Reconstruction Residuals.
Iterative soft thresholding reconstruction ℓ_2 residuals of (**A**) 192 ^{1}H–^{15}N HSQC F_1 traces and (**B**) 10 HNCA $F_2 - F_1$ planes from Poisson-gap schedules having sampling densities of 30% (blue), 10% (green) and 5% (red). Residuals of sine-gap and sine-burst schedules are shown as solid and dashed vertical lines, respectively.

constructed schedule by setting sampled grid points to one and skipped grid points to zero. At each sampling density and grid size for which schedules were created, point-spread functions were calculated by hypercomplex discrete Fourier transformation of each schedule's impulse set. Point-spread functions for schedules built on two-dimensional grids are shown for each sampling density in Figure 2.5. For one-dimensional schedules, reconstruction residuals were computed from a subset of 192 F_1 traces of the half-transformed HSQC spectrum. The traces were nonuniformly subsampled using

sine-gap, sine-burst and Poisson-gap (N = 10,000) schedules and reconstructed with 400 iterations of IST at a threshold level of 98%. After reconstruction, the residual was calculated using the ℓ_2-norm of the differences between the true and reconstructed signals. A convergence analysis was also performed (cf. Figure 2.8) to ensure convergence of IST to a stationary point, as measured by a lack of decrease in the ℓ_2 error. Figure 2.9A shows the distributions of IST reconstruction residuals from the HSQC traces. Reconstructions of 10 $F_2 - F_1$ planes of the F_3-transformed HNCA were also performed after nonuniformly subsampling using sine-gap schedules, sine-burst schedules, and a subset (N = 10,000) of the generated Poisson-gap schedules. Figure 2.9B shows IST reconstruction residuals computed from the HNCA planes, and example reconstructions from each sampling schedule at 5% density are illustrated in Figure 2.7.

Table 2.1: Peak-picking Performances from IST-reconstructed HSQC Spectra.

Method		**Matched**	**Lost**	**Gained**	$\boldsymbol{\rho}$	$\mathbf{d_H}$	$\mathbf{d_N}$
PG	30%	99/99	0/99	2	0.9994	0.000724	0.004459
	10%	99/99	0/99	4	0.9983	0.001208	0.008316
	5%	98/99	1/99	8	0.9920	0.001430	0.009398
SG	30%	99/99	0/99	0	0.9996	0.000580	0.005957
	10%	98/99	1/99	6	0.9983	0.001546	0.007809
	5%	98/99	1/99	7	0.9939	0.001660	0.011393
SB	30%	99/99	0/99	1	0.9996	0.000534	0.008977
	10%	98/99	1/99	5	0.9981	0.001071	0.010007
	5%	98/99	1/99	7	0.9699	0.001482	0.013357

Table 2.2: Peak-picking Performances from IST-reconstructed HNCA Spectra.

Method		**Matched**	**Lost**	**Gained**	$\boldsymbol{\rho}$	$\mathbf{d_H}$	$\mathbf{d_N}$
PG	30%	73/74	1/74	0	0.9978	0.000532	0.007556
	10%	70/74	4/74	0	0.9905	0.001176	0.015378
	5%	66/74	8/74	0	0.9745	0.001488	0.015092
SG	30%	73/74	1/74	0	0.9955	0.000585	0.010554
	10%	66/74	8/74	1	0.9864	0.001793	0.016878
	5%	64/74	10/74	0	0.9638	0.001903	0.020252
SB	30%	73/74	1/74	0	0.9977	0.000560	0.010475
	10%	69/74	5/74	0	0.9883	0.001311	0.015739
	5%	66/74	8/74	1	0.9781	0.001852	0.017306

2.3.5 Generation of Peak-picking Statistics

A summary of the relative HSQC peak-picking performance for the IST reconstructions from each sampling schedule and at each sampling density is listed in Table 2.1. For each 2D ^{1}H–^{15}N HSQC

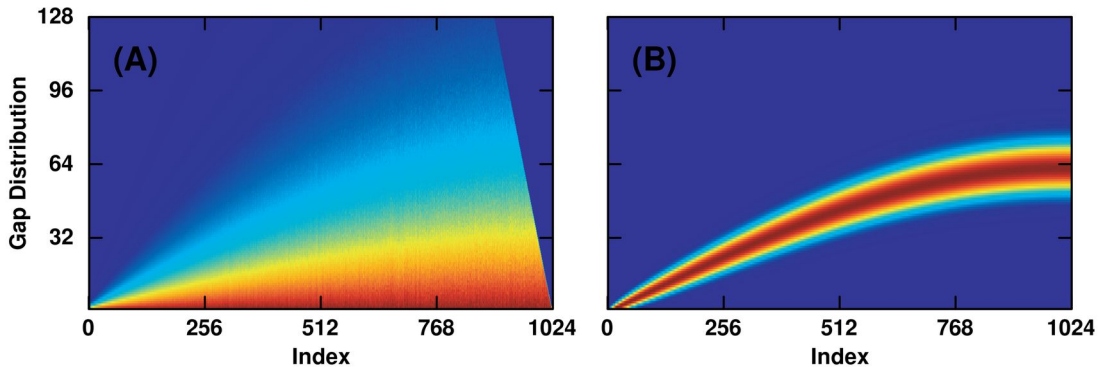

Figure 2.10: Gap Length Histograms from Rejection Sampling.
(**A**) Normalized histograms of gap lengths at various points in an ensemble of 10,000,000 sampling schedules generated by rejection sampling from the Poisson-gap expectation sampling distribution (equation 2.11). (**B**) Normalized histograms of gap lengths produced by true Poisson-gap sampling. Examination of these histograms clearly indicates that the schedules generated by rejection sampling are *not* Poisson-gap schedules.

spectrum of ubiquitin reconstructed via IST at each sampling density and each sampling method, a set of quality statistics was computed. Peak lists were generated using the `peakHN.tcl` utility provided by NMRPipe [3], with a minimum intensity threshold of 3.0×10^7. Then, a greedy algorithm was used to generate a maximum-cardinality bipartite matching between the peak list of each reconstructed spectrum and the peak list of the true spectrum. Chemical shift windows of 0.015 ppm and 0.08 ppm were used along the ^{1}H and ^{15}N dimensions, respectively, during matching. The numbers of peaks matched, lost and gained in the reconstructed spectra, relative to the true spectrum, were all counted. Lost peaks were any picked peaks in the true spectrum that had no match in the reconstruction. Gained peaks were any picked peaks in the reconstruction with no partner in the true spectrum. The intensities of all matched peaks in each reconstruction were then compared against their true intensities through the computation of a Pearson correlation coefficient, ρ, which effectively summarizes the linearity of the reconstruction algorithm as a function of sampling schedule. Finally, root-mean-square chemical shift deviations of all matched peaks along the ^{1}H dimension (d_H) and the ^{15}N dimension (d_N) were also computed. Identical procedures and parameters, with the exception of an intensity threshold of 6.0×10^8, were used to peak-pick the ^{1}H–^{15}N projections of the uniform and reconstructed HNCA spectra (cf. Table 2.2).

2.3.6 Analysis of Sampling Distributions

Expectation sampling distributions were also generated from the set of Poisson-gap schedules by averaging their resulting impulse sets. Figure 2.2 shows the expectation sampling distributions for one-dimensional schedules having different sampling densities, and Figures 2.3 and 2.4 show the distributions for two-dimensional schedules having the same densities. The heavy bias towards early time points in Poisson-gap sampling is reaffirmed in all figures. Sampling distributions were also computed via equations 2.11 and 2.12 for comparison to the distributions obtained by averaging multiple impulse sets (Figures 2.2 and 2.3). To verify that fully random sampling from equation 2.11 and gap sampling from g_{PG} are not equivalent, 10,000,000 sampling schedules were generated by rejection sampling 51 grid points from equation 2.11 at $\Lambda = 62.9$ and $N = 1024$, and histograms of the gap lengths at each grid point were computed (Figure 2.10). If the two methods were indeed equivalent, one would expect the histograms in Figure 2.10A to resemble Poisson distributions (2.10B).

2.3.7 Average Poisson-gap Sequences

For Figure 2.1, each schedule $\{x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}\}$ in the generated ensemble of M (here, $M =$ 50,000) one-dimensional Poisson-gap schedules was averaged on a term-by-term basis:

$$\langle x_i \rangle = \frac{1}{M} \sum_{m=1}^{M} x_i^{(m)} \tag{2.13}$$

to produce the average Poisson-gap schedule $\{\langle x_1 \rangle, \langle x_2 \rangle, \ldots, \langle x_n \rangle\}$. Similar procedures were performed to compute the standard deviation of the Poisson-gap ensemble. Deterministic sampling schedules with a 1x exponential bias were computed according to procedures outlined by Eddy et al. [4]. Relative errors (Figure 2.1) between a given sine-gap schedule $\{y_1, y_2, \ldots, y_n\}$ and average Poisson-gap schedule were computed by a term-by-term subtraction of one schedule from the other, followed by a division by the average Poisson-gap terms:

$$\Delta_i = \frac{y_i - \langle x_i \rangle}{\langle x_i \rangle} \quad \forall i \in \{1, 2, \ldots, n\} \tag{2.14}$$

Relative errors between the deterministic exponential schedules and the average Poisson-gap sequence were similarly computed.

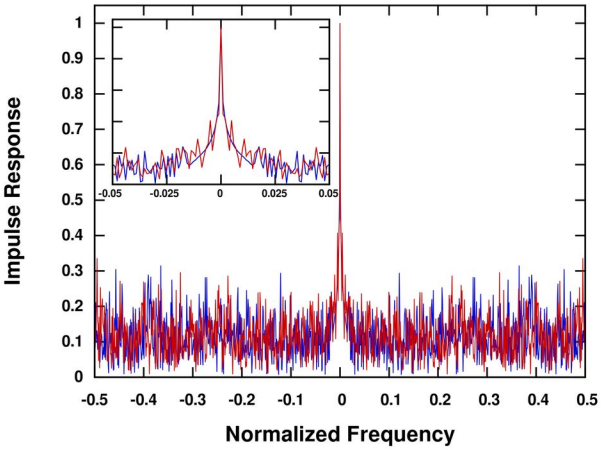

Figure 2.11: Introduction of Spurs by Squared-sine Modulation. Impulse response functions (IRFs) of sine-gap (blue) and sine-burst (red) sampling schedules at 5% density, demonstrating the appearance of low-frequency spurs induced by burst augmentation of the gap equation.

2.4 Results

While at first glance, the deterministic schedule constructed using g_{SG} in Figure 2.5 may appear unrelated to the Poisson-gap schedule, it is in fact a realization of Poisson-gap sampling in which all random draws from the underlying Poisson distribution have resulted in the expected value. This fact is corroborated by the corresponding point-spread functions, which closely resemble those of the stochastic example at 30% and 10% sampling density. Reconstruction residuals from IST (Figure 2.7) also reveal a high similarity between the deterministic sine-gap and stochastic Poisson-gap schedules at 30% and 10% density. However, the sine-gap PSF becomes less comparable to that of Poisson-gap at low sampling densities, where the benefits of incoherent sampling are more apparent. It is worth noting that the striking appearance of sampling artifacts in the sine-gap PSF is a consequence of the log-scaled color gradient used in Figure 2.5, which was necessary in order to visually expose very low-intensity artifacts.

The addition of burst augmentation in the form of g_{SB} does not substantially alter IST reconstruction residuals relative to g_{SG} and g_{PG}. However, artifacts arising from regularity in g_{SG}-based schedules at low sampling densities are diminished by burst augmentation, resulting in point-spread functions that more closely resemble those from stochastic Poisson-gap sampling. This reduction of artifacts by burst augmentation comes at a small cost, at low-frequency spurs are introduced into the sine-burst point-spread function (Figure 2.11) by modulating the gap equation. However, these spurs are low in magnitude and only readily apparent at very low (5%) sampling density. These spurs could potentially be reduced by burst-modulating each dimension in the schedule by a different factor.

IST residuals of sine-burst schedules (Figure 2.9, dashed lines) are slightly greater than those of one-dimensional sine-gap schedules and dense two-dimensional sine-gap schedules, but they improve relative to sine-gap as sampling density is decreased. Therefore, while sine-gap sampling is a valuable tool for understanding the nature of Poisson-gap sampling, it is clearly bested in performance by multidimensional sine-burst sampling as global sampling density is decreased. Burst augmentation re-introduces sampling incoherence into highly coherent sine-gap schedules to produce sine-burst schedules that more closely resemble Poisson-gap sampling schedules. This added incoherence is clearly evident in the ^{1}H–^{13}C projections of reconstructed HNCA spectra (Figure 2.7), where the more incoherent sine-burst schedule yields a more faithful spectral reconstruction than the sine-gap schedule can.

2.5 Discussion and Conclusions

This chapter has shown that Poisson-gap sampling is a single instance in a class of gap sampling methods, which may or may not be defined stochastically. Using the well-defined gap sampling algorithm, two novel deterministic sampling methods have been described: sine-gap and sine-burst sampling. Neither of these new methods relies on random deviates, and both have comparable performance to Poisson-gap sampling according to IST reconstruction residuals. From a practical perspective, Poisson-gap, sine-gap and sine-burst sampling methods produced nearly equivalent HSQC spectral reconstructions (Figure 2.6) that yielded essentially identical information (chemical shifts, peak intensities) as highlighted in Table 2.1. Poisson-gap and sine-burst sampling also produced nearly equivalent HNCA spectra (Figure 2.7) after IST reconstruction, even at low sampling density. Table 2.2 also summarizes the peak-picking statistics collected on ^{1}H–^{15}N projections of the reconstructed HNCA spectra. For the practicing spectroscopist, this equates to the ability to nonuniformly sample at the performance level of Poisson-gap, without specifying a pseudorandom seed. Gap sampling is a flexible and attractive alternative to traditional probabilistic sampling methods that use probability densities to define the local sampling density over a Nyquist grid. In effect, gap sampling approaches the problem of local sampling density from the opposite direction of probabilistic sampling by defining the distances *between* samples on the grid. This chapter also holds a brief derivation of the mathematical connection between stochastic gap equations and their expectation sampling distributions, which allows for direct visualization of the grid-point weighting produced by a given gap equation. While these expectation sampling distributions are useful in describing the sampling behavior of a stochastic gap equation, they do not provide a means of con-

verting a gap-based sampling method into a fully random sampling method. In other words, it has been shown that any method of constrained random sampling using a gap equation is inequivalent to fully random sampling from its corresponding expectation sampling distribution.

Finally, burst augmentation provides a concrete example of how deterministic gap sampling may be tuned to behave in a similar fashion to pseudorandom numbers. At first glance, the third row of Figure 2.5 would appear to have been generated stochastically, but it is a consequence of the squared-sine modulation term in g_{SB}. It has historically been true that stochastically generated sampling schedules produced fewer prominent artifacts than deterministic methods such as radial or spiral sampling, due to high regularity (i.e. coherence) of the latter schemes. However, burst augmentation demonstrates that pseudorandom variates are not strictly required for producing incoherent sampling methods. Furthermore, while most pseudorandom number generators are indeed deterministic for a given seed value, this determinism is *inherently different* from the determinism offered by sine-gap and sine-burst sampling. By design, any parameters (e.g. reconstruction residuals) measured from pseudorandomly generated sampling schedules will not be smoothly varying – and therefore optimizable – functions of their random seed value. As a consequence, no absolute guarantee of spectral quality is provided to the spectroscopist employing pseudorandom sampling schedules, even if the relative difference in quality between the best- and worst-performing Poisson-gap seed values is small at sampling densities above 30%. This problem with seeds has already been recognized: Poisson-gap and jittered sampling methods are, in fact, two separate attempts at minimizing – but not removing – the effect of seed values on schedule performance [11, 14, 17]. Deterministic gap sampling completely frees the user from specifying an arbitrary seed value, and provides a highly general framework that enables further investigation into which features of NUS schedules yield higher-quality reconstruction results.

The C implementations of Poisson-gap, sine-gap and sine-burst sampling are free and open source software, and are available for download at `http://bionmr.unl.edu/dgs.php`. The programs are highly portable and C99 compliant, so they may be compiled on any modern operating system. An online schedule generation tool is also provided at the same address for rapid generation of one-, two- and three-dimensional NUS schedules suitable for direct use on Bruker or Agilent spectrometers. As defined and implemented, the recursive schedule generation algorithm is not limited to any number of grid dimensions. However, the online tool has been hard-limited to 3D grids to minimize server load.

2.6 References

[1] P. C. Aoto, B. Fenwick, G. J. A. Kroon, and P. E. Wright. Accurate scoring of non-uniform sampling schemes for quantitative NMR. *Journal of Magnetic Resonance*, 246:31–35, 2014.

[2] G. L. Bretthorst. Nonuniform Sampling: Bandwidth and Aliasing. *Concepts in Magnetic Resonance*, 32A(6):417–435, 2008.

[3] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe – a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular NMR*, 6(3):277–293, 1995.

[4] M. T. Eddy, D. Ruben, R. G. Griffin, and J. Herzfeld. Deterministic schedules for robust and reproducible non-uniform sampling in multidimensional NMR. *Journal of Magnetic Resonance*, 214:296–301, 2012.

[5] J. C. Hoch, M. W. Maciejewski, and B. Filipovic. Randomization improves sparse sampling in multidimensional NMR. *Journal of Magnetic Resonance*, 193(2):317–320, 2008.

[6] J. C. Hoch, M. W. Maciejewski, M. Mobli, A. D. Schuyler, and A. S. Stern. Nonuniform Sampling and Maximum Entropy Reconstruction in Multidimensional NMR. *Accounts of Chemical Research*, 47(2):708–717, 2014.

[7] S. G. Hyberts, H. Arthanari, S. A. Robson, and G. Wagner. Perspectives in magnetic resonance: NMR in the post-FFT era. *Journal of Magnetic Resonance*, 241:60–73, 2014.

[8] S. G. Hyberts, H. Arthanari, and G. Wagner. Applications of Non-uniform Sampling and Processing. *Topics in Current Chemistry*, 316:125–148, 2012.

[9] S. G. Hyberts, A. G. Milbradt, A. B. Wagner, H. Arthanari, and G. Wagner. Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. *Journal of Biomolecular NMR*, 52(4):315–327, 2012.

[10] S. G. Hyberts, S. A. Robson, and G. Wagner. Exploring signal-to-noise ratio and sensitivity in non-uniformly sampled multi-dimensional NMR spectra. *Journal of Biomolecular NMR*, 55(2):167–178, 2013.

[11] S. G. Hyberts, K. Takeuchi, and G. Wagner. Poisson-Gap Sampling and Forward Maximum Entropy Reconstruction for Enhancing the Resolution and Sensitivity of Protein NMR Data. *Journal of the American Chemical Society*, 132(7):2145–2147, 2010.

[12] L. E. Kay, P. Keifer, and T. Saarinen. Pure Absorption Gradient Enhanced Heteronuclear Single Quantum Correlation Spectroscopy with Improved Sensitivity. *Journal of the American Chemical Society*, 114(26):10663–10665, 1992.

[13] K. Kazimierczuk and V. Y. Orekhov. Accelerated NMR spectroscopy by using compressed sensing. *Angewandte Chemie*, 50(24):5556–5559, 2011.

[14] K. Kazimierczuk, A. Zawadzka, W. Kozminski, and I. Zhukov. Lineshapes and artifacts in Multidimensional Fourier Transform of arbitrary sampled NMR data sets. *Journal of Magnetic Resonance*, 188(2):344–356, 2007.

[15] M. W. Maciejewski, M. Mobli, A. D. Schuyler, A. S. Stern, and J. C. Hoch. Data Sampling in Multidimensional NMR: Fundamentals and Strategies. *Topics in Current Chemistry*, 316:49–78, 2012.

[16] M. W. Maciejewski, H. Z. Qui, I. Rujan, M. Mobli, and J. C. Hoch. Nonuniform sampling and spectral aliasing. *Journal of Magnetic Resonance*, 199(1):88–93, 2009.

[17] M. Mobli. Reducing seed-dependent variability of non-uniformly sampled multidimensional NMR data. *Journal of Magnetic Resonance*, 256:60–69, 2015.

[18] M. Mobli and J. C. Hoch. Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 83C:21–41, 2014.

[19] A. G. Palmer, J. Cavanagh, P. E. Wright, and M. Rance. Sensitivity Improvement in Proton-Detected Two-Dimensional Heteronuclear Correlation NMR-Spectroscopy. *Journal of Magnetic Resonance*, 93(1):151–170, 1991.

[20] M. R. Palmer, B. R. Wenrich, P. Stahlfeld, and D. Rovnyak. Performance tuning non-uniform sampling for sensitivity enhancement of signal-limited biological NMR. *Journal of Biomolecular NMR*, 58(4):303–314, 2014.

[21] D. Rovnyak, D. P. Frueh, M. Sastry, Z. Y. J. Sun, A. S. Stern, J. C. Hoch, and G. Wagner. Accelerated acquisition of high resolution triple-resonance spectra using non-uniform sampling and maximum entropy reconstruction. *Journal of Magnetic Resonance*, 170(1):15–21, 2004.

[22] D. Rovnyak, J. C. Hoch, A. S. Stern, and G. Wagner. Resolution and sensitivity of high field nuclear magnetic resonance spectroscopy. *Journal of Biomolecular NMR*, 30(1):1–10, 2004.

[23] A. D. Schuyler, M. W. Maciejewski, H. Arthanari, and J. C. Hoch. Knowledge-based nonuniform sampling in multidimensional NMR. *Journal of Biomolecular NMR*, 50:247–262, 2011.

[24] A. D. Schuyler, M. W. Maciejewski, A. S. Stern, and J. C. Hoch. Formalism for hypercomplex multidimensional NMR employing partial-component subsampling. *Journal of Magnetic Resonance*, 227:20–24, 2013.

[25] A. S. Stern, D. L. Donoho, and J. C. Hoch. NMR data processing using iterative thresholding and minimum ℓ_1-norm reconstruction. *Journal of Magnetic Resonance*, 188:295–300, 2007.

[26] S. J. Sun, M. Gill, Y. F. Li, M. Huang, and R. A. Byrd. Efficient and generalized processing of multidimensional NUS NMR data: the NESTA algorithm and comparison of regularization terms. *Journal of Biomolecular NMR*, 62:105–117, 2015.

Chapter 3
Multivariate Analysis in Metabolomics

Essentially, all models are wrong, but some are useful.

– George E. P. Box

3.1 Introduction

The applications of chemometrics are as broad as the field of chemistry itself, but one particularly challenging subdiscipline of bioanalytical chemistry – known as "metabolomics" – has recently renewed interest in the use of chemometrics [107]. Indeed, the chemical complexity of systems studied by metabolomics *necessitates* the use of chemometric techniques: if metabolomics were a nail, chemometrics would surely be a hammer.

Metabolomics is defined [59] as "the quantitative measurement of the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification." Such a definition implies that metabolomics studies offer the finest-grained detail available in the nascent field of systems biology: a molecular-level convolution of all upstream genomic, transcriptomic and proteomic responses of an organism to a given stimulus or change [52, 87, 95]. Metabolites are the end product of all cellular processes, and their in vivo concentrations are a direct result of enzymatic activity. While a change in the expression level of a protein or its coding gene may not necessarily correlate directly with the activity of that protein, alterations in metabolite concentrations *are* the consequence of altered activity [84]. Thus, metabolites are more proximal to a phenotype or disease state than either genetic or proteomic information. The richness of phenotypic information offered by metabolomics has been leveraged to identify disease biomarkers [41, 93], to aid in the drug discovery process [69, 101], and to study plants [45], bacteria [111, 82], nutrition [66], and the environment [12], among numerous other applications [3].

The rich information promised by metabolomics does not come without a price, and metabolomics experiments are plagued with difficulty. The number of small-molecule metabolites in a biofluid,

cell lysate, tissue or organ differs wildly depending on the organism studied, ranging from several hundred to hundreds of thousands [32]. While databases of commonly encountered metabolites have been compiled [102, 24, 53], they are by no means complete. Therefore, it is common to encounter unknown signals during data analysis, complicating the interpretation of metabolic changes between experimental groups. Metabolite identification is further complicated by a lack of NMR or mass spectral reference information for known metabolites. Finally, the diversity of chemical and physical properties of metabolites makes true simultaneous quantitation of all metabolites present in a system unattainable with current instrumental capabilities [59, 32, 29]. As an illustration, due to the limited molecular mass distribution of the metabolome, comprehensive metabolomic analyses by mass spectrometry generally require the prefixing of one or more chromatographic separations prior to analyte ionization [52, 94].

The extraction of information from data in metabolomics experiments is further complicated by the inherent variability present within each sample. Every single cell, tissue, organ or organism is fundamentally unique [71], despite any features (disease state, drug treatment, etc.) it may have in common with others of its kind. Thus, the differentiation between two experimental groups in a metabolomics experiment requires the identification of relatively few defining or discriminating chemical features against a large, complex background of metabolites [102]. Ideally, these few chemical features may be identified as a unique set of metabolites that are directly related to the defining biochemical states of each experimental group. Unfortunately, all biological systems are easily perturbed by experimental or environmental factors, including age, gender, diet, cell growth phase, nutrient availability, pH and temperature [90, 111]. Variations in sample handling procedures, including cell lysis, metabolic quenching, metabolite extraction and sample storage can also introduce further variability into the measured data. Finally, variations in signal position, intensity and shape may manifest from instrumental instabilities on a per-sample basis. Each of these numerous sources of sample variability increases the magnitude of $\mathbf{E}$ in any chemometric model that may be applied to the data (cf. Section 1.1), which decreases the statistical validity of $f(\mathbf{D})$. Therefore, the design of experiments and analysis of data in metabolomics requires robust methodologies in order to expose underlying chemical trends from highly complex systems in the form of statistically valid mathematical models. This chapter describes the theory and best practices of chemometric analyses of data produced by metabolomics experiments, with a focus on 1D ^{1}H and 2D ^{1}H–^{13}C NMR datasets.

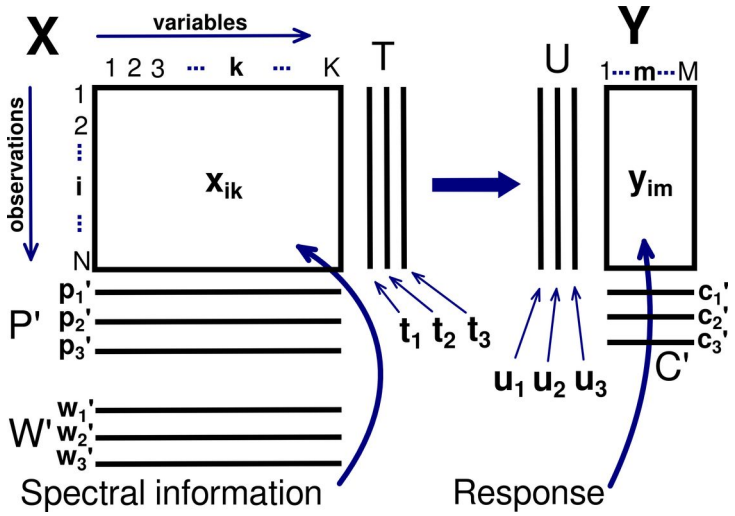

Figure 3.1: Canonical Example of a Bilinear Modeling Problem.
Illustration of a data matrix $\mathbf{X}$ and a response matrix $\mathbf{Y}$, as they are typically used in partial least squares modeling problems. In metabolomics applications, the data matrix will contain a set of N spectra, each having K variables. For supervised modeling problems, each observation in the data matrix is paired with a corresponding row in the response matrix that holds either continuously varying outputs or binary class memberships. The data are then decomposed into a small number of score vectors ($\mathbf{t}$) and loading vectors ($\mathbf{p}$), with corresponding weight vectors ($\mathbf{w}$) used to transform the observations in $\mathbf{X}$ into scores-space. The responses in $\mathbf{Y}$ are similarly decomposed into scores ($\mathbf{u}$) and loadings ($\mathbf{c}$). Tick marks denote transposition.

3.2 Multivariate Datasets

In the majority of cases, multivariate datasets used in metabolomics take the form of second-order tensors in $\mathbb{R}^{N\times K}$. More simply, these datasets are real matrices having N rows and K columns. By convention, the data are arranged as N observation row vectors of length K, where K is referred to as the dimensionality of the dataset (Figure 3.1). Typical examples of 1D datasets include sets of ^{1}H or ^{13}C NMR spectra [4, 55], direct-injection mass spectra (DI-MS, [13, 79, 112]), infrared (IR) and Raman spectra [35, 16], or capillary electrophoretograms (CE, [70]). This remarkable diversity of instrumental platforms used in metabolomics is traceable to the ability of bilinear factorizations such as principal component analysis (PCA, [51]) and partial least squares (PLS, [103]) to directly accept these second-order tensors for modeling (vide infra).

The dimensionality of a multivariate dataset may be increased by adding another "mode", resulting in a third-order (or higher) tensor ($\mathbf{X} \in \mathbb{R}^{N\times K_1\times K_2}$, Figure 3.2). In such cases, the total dimensionality of the dataset is now the product of the dimensionalities along each mode of the data tensor (e.g. $K_1 \times K_2$). Third-order tensors are the natural data structures for sets of two-dimensional ob-

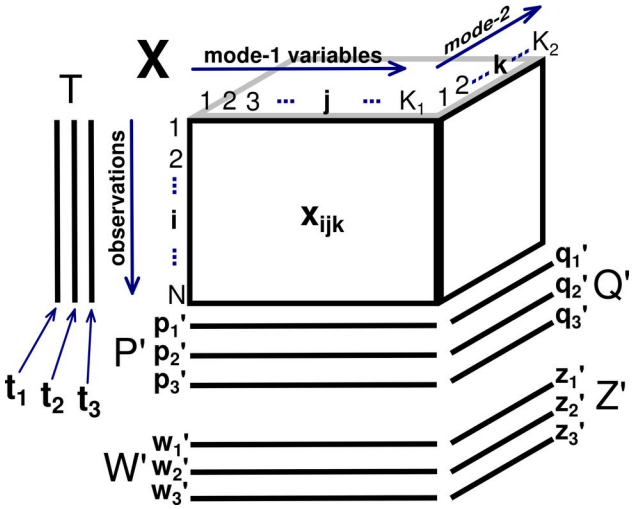

Figure 3.2: Canonical Example of an Unsupervised Trilinear Modeling Problem.
Illustration of a third-order data tensor $\mathbf{X}$ as may be found in multilinear factorization problems. Such data tensors will contain a set of N spectra, each having K_1 variables along their first mode and K_2 variables along their second. The data tensor is then decomposed into a small number of score vectors ($\mathbf{t}$) and loading vectors ($\mathbf{p}$, $\mathbf{q}$), with corresponding weight vectors ($\mathbf{w}$, $\mathbf{z}$) used to transform the observations in $\mathbf{X}$ into scores-space. Tick marks denote transposition.

servations, including ^{1}H–^{1}H, ^{1}H–^{13}C and ^{1}H–^{15}N NMR spectra, hyphenated chromatography-mass spectra (LC-MS, GC-MS), hyphenated electrophoresis-mass spectra (CE-MS), and hyphenated ion-mobility mass spectra (IM-MS). While third-order data tensors may hold substantially more chemical information than their second-order counterparts, they are not directly compatible with bilinear factorization methods, and they require specialized processing, treatment and modeling algorithms [62, 63]. As an example, tensors may be vectorized into matrices [46] that are suited for PCA and PLS, but at the cost of lost structural information. Methods such as uncorrelated multilinear PCA (UMPCA, [62]), on the other hand, provide a means of directly decomposing tensors into low-dimensional spaces while maintaining structural information.

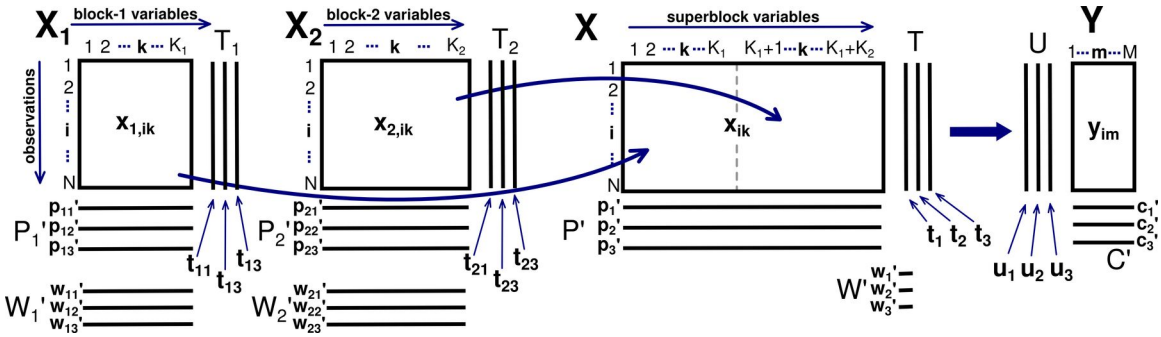

Figure 3.3: Canonical Example of a Multiblock Bilinear Modeling Problem.
Illustration of a pair of data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, their observation-wise concatenation $\mathbf{X}$, and a response matrix $\mathbf{Y}$, as they are typically used in multiblock partial least squares modeling problems. In metabolomics applications, each data matrix $\mathbf{X}_b$ in the set of B matrices will contain a set of N spectra, each having K_b variables. The data are then decomposed into a small number of superblock score vectors ($\mathbf{t}$) and superblock loading vectors ($\mathbf{p}$), with corresponding superblock weight vectors ($\mathbf{w}$). Each individual data matrix is also decomposed into a set of block scores ($\mathbf{t}_b$), block loadings ($\mathbf{p}_b$) and block weights ($\mathbf{w}_b$). Tick marks denote transposition.

Another mechanism of increasing the information content of datasets entering into multivariate chemometric models is to collect spectral observations on two or more complementary instrumental platforms for each sample. In such a "multiblock" modeling approach, each data block $\mathbf{X}_b$ contains N K_b-variate observations [99, 77]. Bilinear methods such as consensus PCA (CPCA), hierarchical PCA and PLS, and multiblock PLS may then be used to provide information about data variation that is correlated between data blocks. Recent examples of multiblock modeling in metabolomics include fusions of near-IR and mid-IR spectra [6], ^{1}H NMR and direct injection electrospray mass spectra (DI-ESI-MS, [65]), and observations from multiple sensors in process control applications [39].

3.3 Spectral Processing

Following the acquisition of experimental data, instrumentation-specific processing must be applied to transform the data into a suitable set of real matrices for bilinear modeling, or tensors for multilinear modeling. Because the majority of data presented herein originated from an NMR spectrometer, and because NMR spectra present unique challenges to the analyst during data handling, the following discussions will center around processing of 1D and 2D NMR datasets.

3.3.1 NMR Signals

Modern NMR spectrometers effectively acquire a rotating-frame free induction decay (FID) through the use of quadrature phase detection of the incoming signal [57]. This detection method imparts relative phase information to the time-domain decays by creating an "in-phase" signal component $i(t)$ and a "quadrature" component $q(t)$ phased ninety degrees from $i(t)$. Indirect dimensions of multidimensional NMR experiments are also collected in quadrature through interleaved acquisition of one-dimensional decays that have been cosine- and sine-modulated by the indirect-dimension signals [80]. As a result, each data point in a D-dimensional NMR signal collected in complete quadrature exists in a hypercomplex space $\mathbb{H}_D$ [73], which is defined by a real basis element and D complex basis elements:

$$\Phi_D \equiv \{1 \cdot u_1 \cdots u_D\} \tag{3.1}$$

where multiplication by any complex element u_d results in a ninety degree phase shift in dimension d, and the basis elements combine commutatively under multiplication, as follows:

$$u_i u_j = u_j u_i \tag{3.2}$$

$$u_i^2 = -1 \tag{3.3}$$

The basis elements in Φ_D are a generating set for the complete set of components of the hypercomplex space $\mathbb{H}_D$. For example, in three dimensions:

$$\Phi_3 = \{1, u_1, u_2, u_1u_2, u_3, u_1u_3, u_2u_3, u_1u_2u_3\} \tag{3.4}$$

A scalar in $\mathbb{H}_D$ is then expressed as a linear combination of this component set. For the three-dimensional example:

$$x = a + bu_1 + cu_2 + du_1u_2 + eu_3 + fu_1u_3 + gu_2u_3 + hu_1u_2u_3 \tag{3.5}$$

or, more generally and succinctly:

$$x = \sum_{\phi \in \Phi_D} x\{\phi\} \cdot \phi \tag{3.6}$$

where $x\{\phi\}$ denotes the real coefficient of x that scales the basis component ϕ in x. For the above three-dimensional example, the expression $x\{u_1u_3\}$ would evaluate to f. Finally, the expression of hypercomplex tensors is formally accomplished by defining each coefficient as a real tensor of appropriate size, like so:

$$x\{\phi\} \in \mathbb{R}^{k_1 \cdots k_K} \quad \forall \phi \in \Phi_D \tag{3.7}$$

where K is the number of modes of the tensor. The above equation may be compactly written as $\mathbb{H}_D^{k_1 \cdots k_K}$. Any scalar in $\mathbb{H}_D$ – and therefore any data point in a D-dimensional quadrature-complete NMR dataset – shall require 2^D real coefficients in order to be completely determined. While the hypercomplex algebras $\mathbb{H}_0$ and $\mathbb{H}_1$ are isomorphic to the real ($\mathbb{R}$) and complex ($\mathbb{C}$) numbers, respectively, $\mathbb{H}_2$ and $\mathbb{H}_3$ are *not* isomorphic to the quaternions and octonions, as the latter are non-commutative under multiplication. This hypercomplex algebra, introduced for partial-component nonuniform subsampling by Schuyler et al. [73], provides an elegant formalism for expressing and handling NMR data.

Mathematically, 1D NMR free induction decays are described by the following commonly used parametric signal model:

$$f(t) = \sum_{m=1}^{M} \alpha_m \exp\{u_1(\omega_m t + \theta_m) - \rho_m t\} \tag{3.8}$$

where α_m, ω_m, θ_m and ρ_m represent the amplitude, frequency, phase error and decay rate of the m-th damped complex exponential in the model $f(t)$. Using the above formalism for hypercomplex tensors, this signal model is trivially extended to any number of dimensions by multiplying in a modulation term for each dimension:

$$f(\mathbf{t}) = \sum_{m=1}^{M} \alpha_m \prod_{d=1}^{D} \exp\{u_d(\omega_{m,d} t_d + \theta_{m,d}) - \rho_{m,d} t_d\} \tag{3.9}$$

For example, a 2D FID may be modeled as follows:

$$f(t_1, t_2) = \sum_{m=1}^{M} \alpha_m \exp\{u_1(\omega_{m,1} t_1 + \theta_{m,1}) + u_2(\omega_{m,1} t_2 + \theta_{m,2}) - \rho_{m,1} t_1 - \rho_{m,2} t_2\} \tag{3.10}$$

In short, NMR free induction decays may be treated as sums of damped hypercomplex exponentials. While it is possible to directly parameterize $f(\mathbf{t})$ using either maximum likelihood estimation [19, 20, 17] or Bayesian model selection and estimation [7, 8, 9, 18], this chapter will focus on the soft modeling of multiple NMR spectra using bilinear matrix factorizations. However, the above parametric description of NMR data is useful in understanding various processing tasks required by these hypercomplex tensors.

3.3.2 Time-domain Processing

Processing of acquired NMR data is broken into two stages, where time-domain data is manipulated, transformed into the frequency domain, and further processed using frequency-domain functions [48]. The most routinely used time-domain NMR processing function – and the first to be applied during processing – is referred to as apodization, where the free induction decay tensor is multiplied point-wise by a window function $w(\mathbf{t})$ that varies over $\mathbf{t}$. Multiplication by this window function serves several purposes, including noise reduction, resolution enhancement, shaping of individual resonances and removal of $\sin(x)/x$ truncation artifacts in the frequency domain. During apodization, it is also common practice to selectively scale the first collected data point in an attempt to reduce later frequency-domain baseline distortions [81, 34].

Following apodization, one or more dimensions of the time-domain NMR data may be extended with zeros, a process known as zero-filling. Doubling of the number of data points by zero-filling is a well-established method of increasing both the digital resolution and the signal-to-noise ratio (SNR) of an NMR signal, and further zero-filling only achieves a smoother interpolation of signals in the frequency domain [34]. A final use of zero-filling is to augment the size of a given dimension into a power of two, enabling the use of a fast Fourier transform (FFT, [21]) in lieu of the slower discrete Fourier transform (DFT) to move the data into the frequency domain.

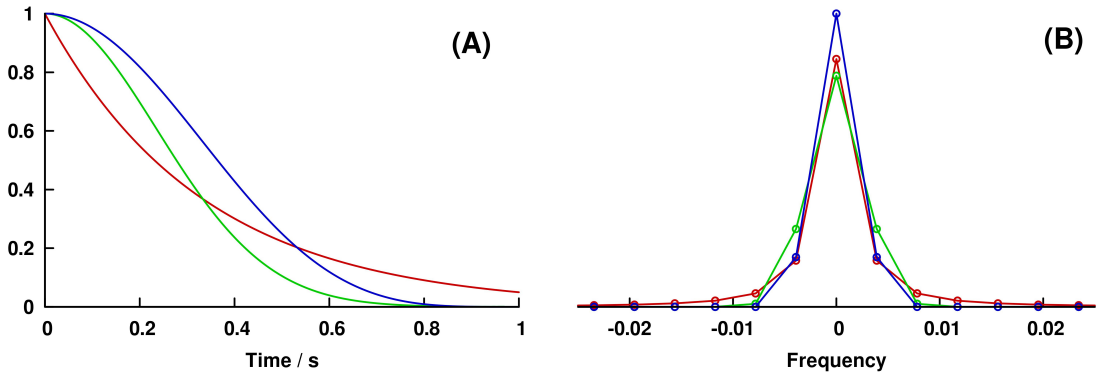

Figure 3.4: Commonly Applied Window Functions.
Window functions (**A**) produced from a 3.0 Hz exponential (red), a 3.0 Hz Gaussian (green), and a squared-cosine (blue). Discrete Fourier transforms of the window functions in (**A**), zoomed around the first few (low frequency) data points, are shown in (**B**). Multiplication of a time-domain signal by a given window function in (**A**) results in a convolution of its frequency-domain counterpart with the impulse response in (**B**). Frequency values in (**B**) are normalized.

3.3.3 Frequency-domain Processing

When NMR free induction decays have been digitized on a grid of uniformly spaced time-domain points, the most convenient method of transforming them into the frequency domain is the discrete Fourier transform (DFT, [10, 73]). Using the introduced formalism for hypercomplex NMR data, the DFT along dimension d of a time-domain vector $\mathbf{f} \in \mathbb{H}_D^N$ is defined[1] as:

$$\mathbf{s}_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \exp\left\{-2\pi u_d \frac{nk}{N}\right\} \mathbf{f}_n \quad \forall k \in \mathbb{Z}_0^{N-1} \tag{3.11}$$

which is a linear transformation $\mathcal{F}_d : \mathbb{H}_D^N \to \mathbb{H}_D^N$. Discrete Fourier transformation of multidimensional NMR data along one dimension requires the application of $\mathcal{F}_d$ to every d-mode vector of the

[1] For succinctness, the set of integers from α to β shall be denoted as $\mathbb{Z}_\alpha^\beta$, i.e. $\mathbb{Z}_\alpha^\beta \equiv \{i \mid i \in \mathbb{Z} \wedge \alpha \leq i \leq \beta\}$

hypercomplex tensor, and full Fourier transformation requires such an operation along each mode of the tensor. Discrete Fourier transformation is computationally efficient when using the FFT, and requires no prior knowledge about the frequency content of the data. However, when only a subset of data points have been collected from a uniform Nyquist grid, as is the case during nonuniform sampling (NUS), the DFT is a sub-optimal estimator of frequency content, and other non-Fourier methods of transformation are required [10, 67].

Once transformed into the frequency domain, NMR spectra require a phase correction processing step, in which a phase factor $\Theta(\omega)$ is multiplied point-wise with the data to correct for phase errors (i.e. $\theta_{m,d}$ terms in $f(\mathbf{t})$) in the data. For example, a 1D phase-factor along dimension d would have the following form:

$$\Theta(\omega) = e^{-u_d \theta(\omega)} \tag{3.12}$$

Ideally, the detected time-domain free induction decays would arrive in-phase with respect to the receiver, and fine tuning of acquisition parameters can often accomplish this [20]. However, variations in receiver phase, dead time between the transmit and receive gating circuits, and delays arising from analog and digital filtering can all introduce phase errors. These phase errors mix the in-phase and quadrature components of the hypercomplex signal, and produce a mixture of desirable absorptive spectral lines and broad dispersive lines between the real and imaginary components of each data point. Unmixing of these absorptive and dispersive contributions to the real spectral component involves the identification of the phase error $\theta(\omega)$, an expansion of phase error terms as powers of ω:

$$\theta(\omega) = \theta_0 + \theta_1\omega + \theta_2\omega^2 + \dots \tag{3.13}$$

Realistically, phase errors higher than first-order are not observed in modern NMR spectra, and phase correction rests on the determination of a zero-order phase error (θ_0) and a first-order phase error (θ_1) in each dimension. This determination may be performed manually, through software-interactive adjustment of zero- and first-order corrections by the analyst. However, manual phase correction is generally too time-consuming in the case of chemometric studies, where there are tens to hundreds of spectra to correct. In that case, the task of phase correction is handed to any number of automated routines that correct each spectrum individually. Spectra may be automatically phase-corrected by maximization of the most negative absorptive data point [75], analysis of the absorption-versus-dispersion [23] or symmetry [47] characteristics of spectral lines, baseline optimization [11]

or entropy minimization [15], to name a few. It is important to note that, when the ultimate fate of the spectral data is multivariate analysis, the phase-correction of each spectrum in isolation is wasteful of information that is available from treating the dataset as an ensemble [108], as phase differences *between* spectra non-linearly affect both line shapes and baseline, possibly emphasizing spectral details that contain no chemically or biochemically relevant information.

3.4 Statistical Treatment

The properties of the bilinear factorizations commonly applied in metabolomics dictate that processed data tensors be treated by one or more operations before they are suitable for modeling. These statistical treatments generally aim to either reduce the dimensionality of the data tensor (i.e. binning and variable selection) or increase the self-consistency of observations and variables (i.e. alignment, normalization and scaling). Treatment operations are usually instrumentation-agnostic, as the data at this stage of handling almost always fall into one of the general structures outlined in Section 3.2.

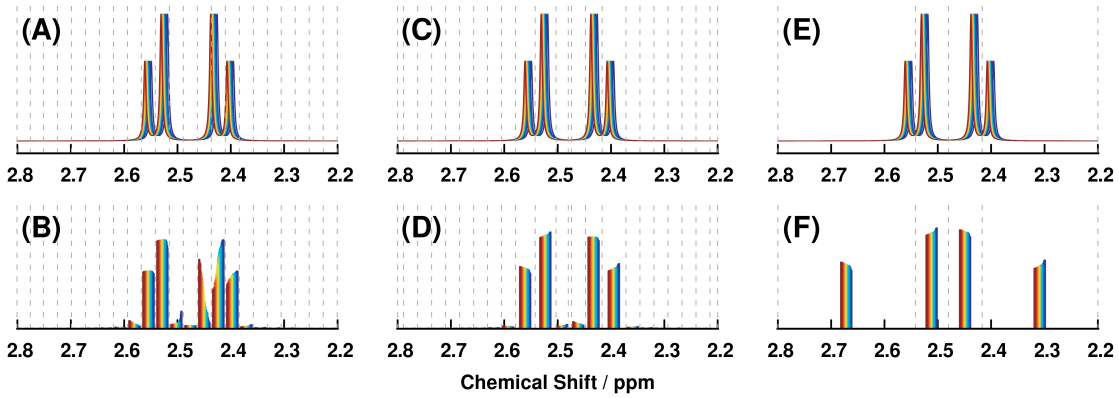

Figure 3.5: Example Bin Region Selection Results.
Simulated ^{1}H NMR spectra of citrate in 20 samples having pH values normally distributed around 6.0 ± 0.05 pH units, binned using uniform (**A**, **B**), optimized (**C**, **D**) and adaptive-intelligent (**E**, **F**) algorithms. Results of bin region integration are shown in the bottom panels.

3.4.1 Binning

Because the chemical shifts of ^{1}H nuclei depend strongly on temperature, pH, ionic strength, and several other factors that affect their electronic environment, spectral datasets acquired for NMR metabolic fingerprinting suffer from imprecision in ^{1}H chemical shifts between observations. This chemical shift imprecision, known as a problem of imperfect correspondence among variables in **X**

[1], decreases the reliability and interpretability of multivariate bilinear models (e.g. PCA, PLS) trained directly on full-resolution spectral data in $\mathbf{X}$. Similar errors in correspondence may also occur in chromatographic datasets, where small drifts in retention time arise from instrumental instability, analyte interactions, and fluctuations in mobile phase and stationary phase composition [68]. The traditional method of mitigating imperfect variable correspondence in a data matrix is to partition the original set of variables into a smaller set of regions, referred to as bins, and to integrate each bin to yield a data matrix having reduced dimensionality.

While binning masks variable mis-correspondence, filters incoherent instrumental noise, and achieves substantial dimensionality reduction, it often hides potentially significant variation in low-intensity resonances nearby strong signals. If bin regions are specified with a uniform size, binning is nearly guaranteed to split signals or spectral features into multiple bins, resulting in undesirable multicollinearities within the reduced variable set. Optimized binning [78] attempts to avoid dividing signals between bins by adjusting uniform bin boundaries into local minima of the data matrix mean, $\langle \mathbf{X} \rangle$. However, because optimized binning begins with a uniform bin set, its practical ability to minimize peak splitting is limited. More complex methods of region identification use either peak detection [25] or recursive subdivision [28] in order to define a more optimal bin set without relying on uniform bin boundaries.

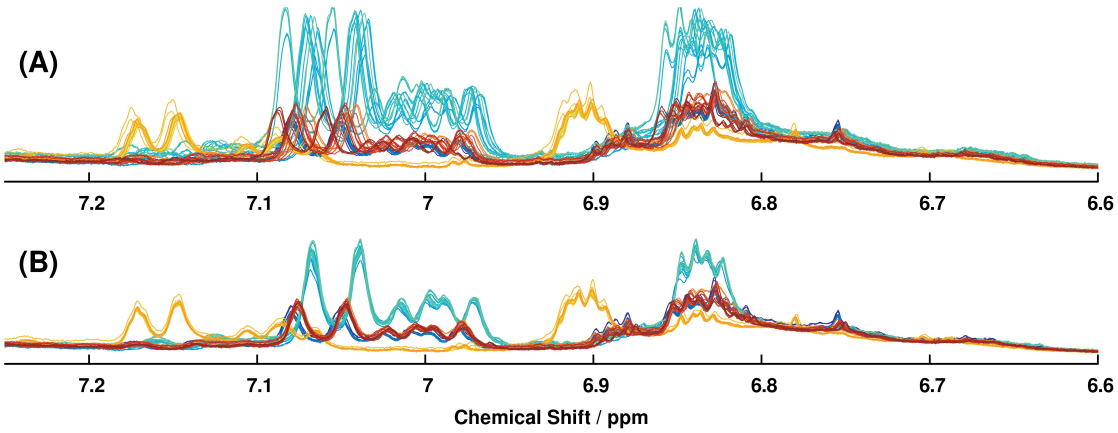

Figure 3.6: Example *i*COshift Alignment Results.
Full-resolution (**A**) and interval correlation-optimized shifting (*i*COshift) aligned (**B**) 1D ^{1}H NMR spectra from a chemometric study of brewed coffee roasts. Spectral alignment was performed such that each observation was shifted to maximize correlation with its respective group mean. Spectral color indicates the observation index.

3.4.2 Alignment

Binning capably masks variability in signal positions and provides an effective means of dimensionality reduction, but it also results in a drastic loss of fine spectral information, as nearby distinct spectral features have been integrated together in the binning process. When full spectral resolution is required during model training and interpretation, the correspondence problem in NMR and chromatographic datasets may be alternatively addressed by signal alignment methods. The most commonly applied alignment algorithms rely on either a warping transformation [68, 40, 109] or linear shifts [92, 72, 85] to bring individual variables of each observation into alignment with a reference observation, which is usually the mean of the data. Warping during alignment is more applicable in situations where a linear dependence between variable index (e.g. retention time) and peak width is expected. In contrast, shift-based alignment preserves peak width, which is ideal for spectroscopic datasets. Like binning, all alignment algorithms must first subdivide the variable set into regions that are then individually warped or shifted, and considerations similar to those in binning apply equally well during alignment region selection.

3.4.3 Normalization

Despite the quantitative nature of most spectroscopic platforms, chemometric samples exhibit variable total analyte concentrations due to variations in sample preparation, instrument stability, or even the samples themselves. These "dilution errors" are especially common in metabolomics experiments using samples of biofluids such as urine, where total concentrations may vary several orders of magnitude. To ensure spectral intensities in a data tensor are directly comparable across each observation, normalization is applied to the tensor [107]. The most common normalization method used in chemometrics is unit-integral or constant-sum (CS) normalization, where each observation is scaled such that its total integral is unity [22]. CS normalization does more harm than good, however, as it introduces false correlations between variables and poorly tolerates large disparities in intensities between each observation.

In an attempt to overcome the drawbacks of CS normalization, Dieterle et al. introduced probabilistic quotient (PQ) normalization, in which the median normalization quotient between all corresponding data points is used as an estimator of the true dilution factor [30]. Shortly after, a method of normalization based on intensity histogram matching (HM) was proposed as an alternative to PQ normalization, taking cues from image processing algorithms [86]. Based on their ability to more ac-

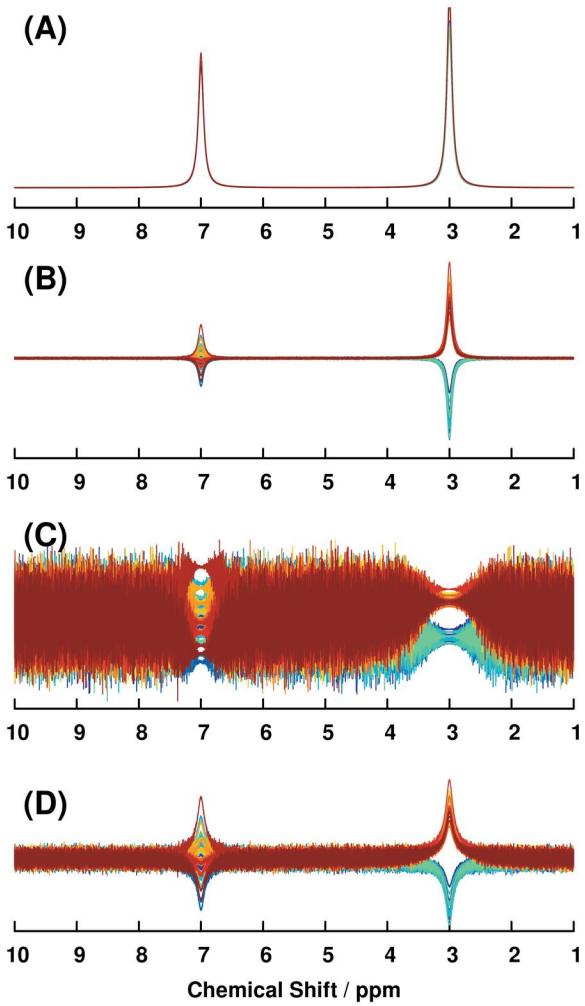

Figure 3.7: Effects of Scaling Noisy Synthetic Spectra.
(**A**) Set of 40 synthetic spectra containing two Lorentzian signals. The first set of signals at 7.0 ppm have normally distributed intensities of 20 ± 0.5 absolute units, and the second set at 3.0 ppm are divided into two sets of intensities, the first normally distributed around 25 ± 0.5 units and the second around 30 ± 0.5 units.
(**B**) The same set of synthetic spectra after subtraction of the sample mean of the data matrix. The two sets of intensities at 3.0 ppm now appear markedly different after centering.
(**C**) The set of synthetic spectra after unit variance (UV) scaling, illustrating the strong noise amplification effect of the UV method.
(**D**) The set of synthetic spectra after Pareto scaling, in which noise amplification is reduced relative to UV scaling.

curately recover true dilution factors, both PQ and HM normalization were reported to outperform CS normalization on real and simulated ^{1}H NMR metabolomics data matrices. Finally, while more commonly applied to IR spectroscopic data, standard normal variate (SNV) and its mathematical cousin, multiplicative scatter correction (MSC) are also candidate methods for normalizing data tensors produced by NMR and other spectroscopic platforms [38].

3.4.4 Scaling

Because bilinear factorization methods such as PCA and PLS generate models based on the eigenstructure of the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ (vide infra), they are sensitive to the magnitudes of individual variables in those matrices. Variables with greater intensity – and therefore greater variance – in a data matrix will draw the attention of these methods, resulting in an unequal weighting of variable importance during model training [51, 76, 91]. As a consequence, analysts commonly apply one or more scaling transformations to their data prior to modeling. The simplest and most

pervasive method, referred to as Unit Variance (UV) scaling, centers each variable with respect to its mean and scales by its standard deviation, like so:

$$\tilde{x}_{nk} = \frac{x_{nk} - \bar{x}_k}{s_k} \tag{3.14}$$

where $\widetilde{\mathbf{X}}$ is the scaled data matrix, $\bar{x}_k$ is the sample mean of the k-th variable over all N observations, and s_k is the corresponding sample standard deviation. Subtraction of the sample mean facilitates the identification of differences between observations, and scaling by the sample standard deviation equally weights every variable in $\mathbf{X}$. When data are UV-scaled, methods that normally analyze covariance eigenstructure will instead rely on *scale-invariant* correlations between variables. Although UV scaling achieves an equal weighting of all variables entering into PCA or PLS, it amplifies the weight of noise variables relative to that of signal variables, resulting in decreased model utility and reliability [49]. Pareto scaling applies a less aggressive scaling than UV by retaining partial covariance between variables in an attempt to reduce this noise amplification:

$$\tilde{x}_{nk} = \frac{x_{nk} - \bar{x}_k}{\sqrt{s_k}} \tag{3.15}$$

A more advanced scaling method that avoids noise amplification uses a maximum likelihood scaling transformation (MALS, [49]) that accounts for the estimated distribution of noise in $\mathbf{X}$. Other forms of scaling have been developed that emphasize various desirable features in a data structure. For example, variable stability (VAST) scaling multiplies each element by the coefficient of variation of its variable in order to focus on highly stable spectral features:

$$\tilde{x}_{nk} = \frac{x_{nk} - \bar{x}_k}{s_k} \cdot \frac{\bar{x}_k}{s_k} \tag{3.16}$$

The alternative level scaling method scales data elements by their sample mean, effectively focusing later analyses on changes in relative magnitude:

$$\tilde{x}_{nk} = \frac{x_{nk} - \bar{x}_k}{\bar{x}_k} \tag{3.17}$$

However, both VAST and level scaling have a more limited scope of application than the general UV, Pareto and MALS methods described above, as they yield optimal transformations only on data structures containing the features they aim to accentuate. For example, VAST scaling is not suited for data tensors that contain large variation between experimental groups, unless further steps are

taken to VAST-scale on a (supervised) per-group basis.

A special case for scaling occurs during multiblock modeling when two or more data blocks contain differing variable counts. In these situations, data blocks having more variables would acquire a larger effective weight during model training. For example, joint modeling of full-resolution ^{1}H NMR data ($K \approx 10^3$) and mass spectral data ($K \approx 10^5$) would result in a weighting of MS variables by a factor of ten relative to NMR variables. To achieve equal block weighting, the variables of each block must be scaled by the square root of the number of block variables:

$$\tilde{\tilde{x}}_{bnk} = \frac{\tilde{x}_{bnk}}{\sqrt{K_b}} \tag{3.18}$$

where the second tilde indicates block scaling in addition to any standard scaling (e.g. UV, Pareto) that may have been applied. When all data blocks contain analyte concentrations instead of raw spectral variables, range scaling may be applied prior to block scaling to remove instrumental response factors and transform all concentrations into relative values [76]:

$$\tilde{x}_{bnk} = \frac{x_{bnk} - \bar{x}_{bk}}{\max_n(x_{bnk}) - \min_n(x_{bnk})} \tag{3.19}$$

Range scaling holds intuitive appeal for multiblock modeling of concentration data, but its application to other kinds of datasets is ill-advised, as it suffers from similar noise amplification problems as UV scaling.

3.4.5 Variable Selection

Due to the expense of sample preparation and data acquisition in metabolomics studies, a strong temptation exists to retain all observed variables during multivariate analyses [54]. Because variables are scaled to equal (or nearly equal) weight prior to modeling, this practice produces multivariate models that suffer in both predictive ability and general reliability. In short, only variables that are truly relevant to the chemical system under study should be included during modeling. To that end, conservative manual removal of irrelevant variables based on spectroscopic and biochemical domain knowledge is often performed in metabolomics. ^{1}H NMR datasets, for instance, nearly always contain highly varying signals from solvents, buffers, and chemical shift reference compounds, all of which may confound or overshadow relevant sources of variation. Variables containing such signals, as well as signal-free variables that only contain instrumental noise, are excellent candidates for

manual variable selection. More computationally intensive methods of variable selection, including support vector machine recursive feature elimination (SVM-RFE), genetic algorithms (GA), random forests (RF) and bootstrapping have also been developed to more aggressively select variables from multivariate data structures [58, 105]. While it is important to retain only relevant variables for modeling, an over-aggressive variable selection is equally detrimental, as it leads to over-fit models that may fail to tolerate subsequent outlying observations.

3.5 Modeling

The most widely applied modeling methods in metabolomics – namely principal component analysis, partial least squares and orthogonal projections to latent structures (OPLS, [88]) – fall within a class of methods known as bilinear matrix factorizations. The general form of a bilinear matrix factorization is:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.20}$$

where the data matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ is approximated by the product of two matrices, $\mathbf{T} \in \mathbb{R}^{N \times A}$ and $\mathbf{P} \in \mathbb{R}^{K \times A}$, which are referred to as "scores" and "loadings", respectively. The matrix $\mathbf{E} \in \mathbb{R}^{N \times K}$ holds any variation in $\mathbf{X}$ that is not captured by the scores and loadings. To understand how such a factorization may be used to approximate a data matrix, it is instructive to consider the product of scores and loadings in vector form:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t_a}\mathbf{p_a}^T + \mathbf{E} \tag{3.21}$$

where $\mathbf{t_a}$ and $\mathbf{p_a}$ are the a-th columns of the score and loading matrices, respectively. In other words, the data matrix is approximated by a set of A rank-1 matrices that are constructed by the outer products of each pair of scores and loadings. Because A is commonly much less than either N or K, these bilinear factorizations are also referred to as low-rank approximations of their data matrices.

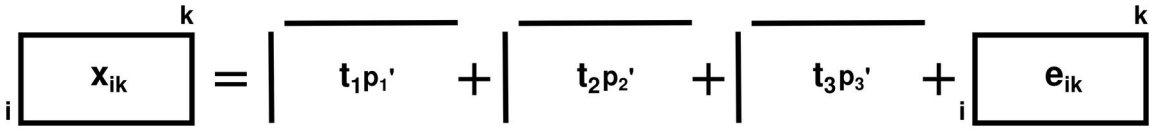

Figure 3.8: Example Three-component Bilinear Low-rank Approximation.
Depiction of a three-component bilinear low-rank approximation $\mathbf{T}\mathbf{P}^T$ of a data matrix $\mathbf{X}$ having residuals $\mathbf{E}$, with each outer product of the approximation broken out.

It is important to note that nearly every data matrix modeled by equation (3.20) in metabolomics contains far fewer observations than variables (i.e. $N \ll K$) [107]. In this situation, there are infinitely many solutions to the equation that yield the same error $\mathbf{E}$. This is easily demonstrated by multiplying the scores and loadings by an orthonormal matrix $\mathbf{R} \in \mathbb{R}^{A \times A}$:

$$\begin{aligned} \mathbf{X} &= \hat{\mathbf{T}}\hat{\mathbf{P}}^T + \mathbf{E} \\ &= \mathbf{T}\mathbf{R}\mathbf{R}^T\mathbf{P}^T + \mathbf{E} \\ &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \end{aligned} \tag{3.22}$$

where $\hat{\mathbf{T}} = \mathbf{T}\mathbf{R}$ and $\hat{\mathbf{P}} = \mathbf{P}\mathbf{R}$. A similar expansion of the solution set may be accomplished by multiplying the scores by a diagonal $A \times A$ matrix, and multiplying the loadings by the inverse of the same diagonal matrix. This equivalence of an infinite number of solutions to equation (3.20), known as the problems of rotational and scale ambiguity, is solved by placing constraints on the values that scores and loadings may take [26, 51]. The choice of constraints defines a particular bilinear factorization method as unique, and determines what kind of information is sought from a data matrix using that method.

3.5.1 Principal Component Analysis

In principal component analysis (PCA), which exactly follows equation 3.20, the loading vectors in $\mathbf{P}$ are constrained to be an orthonormal basis set. More precisely, the loadings produced by PCA are not just any orthonormal basis, but are in fact the first A eigenvectors of the sample covariance matrix $\mathbf{X}^T\mathbf{X}$. In chemometrics, the most commonly used algorithm for constructing PCA models is nonlinear iterative partial least squares (NIPALS, [42]):

Algorithm 3.1 NIPALS Algorithm for PCA

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}$
Output: $\mathbf{t} \in \mathbb{R}^N$, $\mathbf{p} \in \mathbb{R}^K$
1: $\mathbf{t}^{(0)} \sim U_{N \times 1}$ { $\mathbf{t}$ may also be initialized to a column of $\mathbf{X}$ }
2: $k \leftarrow 1$
3: **repeat**
4: $\quad \mathbf{p}^{(k)} \propto \mathbf{X}^T\mathbf{t}^{(k-1)}$
5: $\quad \mathbf{t}^{(k)} \leftarrow \mathbf{X}\mathbf{p}^{(k)}$
6: $\quad k \leftarrow k + 1$
7: **until** $\frac{\|\mathbf{t}^{(k)} - \mathbf{t}^{(k-1)}\|_2}{\|\mathbf{t}^{(k-1)}\|_2} < \varepsilon$

where $\leftarrow$ indicates assignment, and $\propto$ indicates normalized assignment. In others words, the follow-

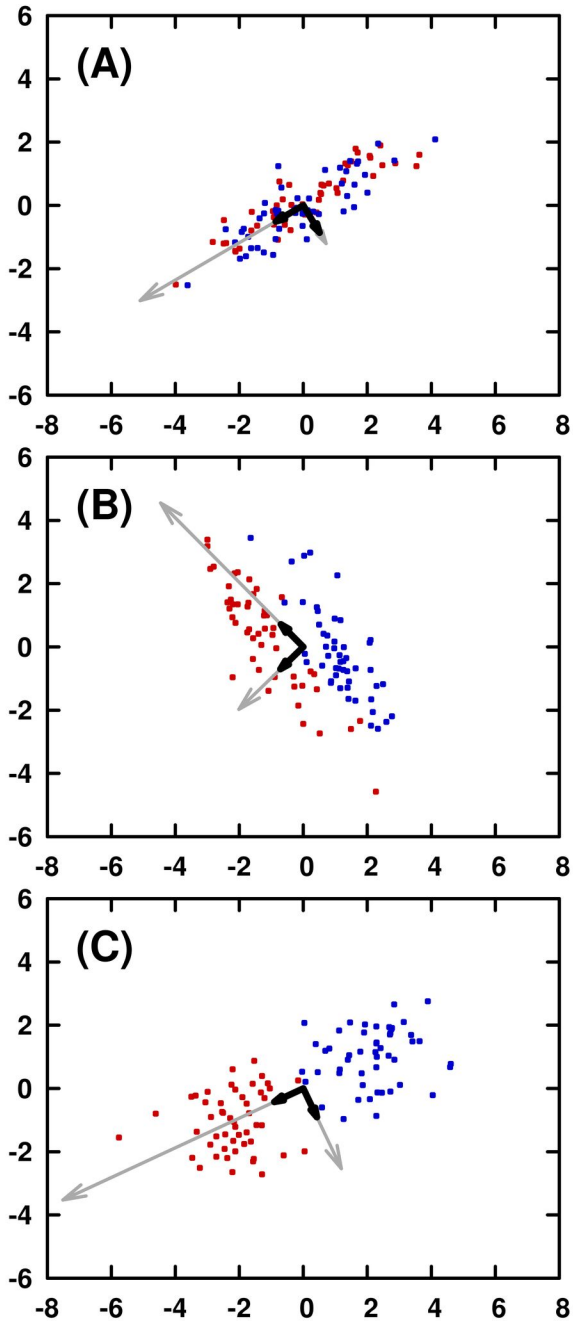

Figure 3.9: Principal Components of Synthetic Bivariate Data.
(**A**) Set of 100 points drawn from a bivariate normal distribution, and the corresponding principal components computed by eigendecomposition of the points' sample covariance matrix. Bold and thin arrows indicate the normalized and unnormalized loadings, respectively.
(**B**) Set of 100 points drawn from two bivariate normal distributions having different means, where the major source of variation in the data is orthogonal to the direction that separates the two groups.
(**C**) Set of 100 points drawn from two bivariate normal distributions having different means, where the major source of variation is parallel to the direction that separates the groups.

ing two statements:

$$a \propto b \quad \Leftrightarrow \quad a \leftarrow \frac{b}{\|b\|_2}$$

are in fact equivalent. In summary, NIPALS PCA initializes a score vector from a uniform random distribution, and repeatedly projects the rows and columns of $\mathbf{X}$ into the score and loading vectors until the scores converge to a predefined limit ε. By substituting the scores assignment into the loadings assignment, we arrive at the following iteration equation:

$$\mathbf{p}^{(k)} \leftarrow \frac{\mathbf{X}^T\mathbf{X}\mathbf{p}^{(k-1)}}{\|\mathbf{X}^T\mathbf{X}\mathbf{p}^{(k-1)}\|_2} \tag{3.23}$$

which is the equation for power iteration on $\mathbf{X}^T\mathbf{X}$ [43]. Indeed, the NIPALS algorithm implicitly computes the dominant eigenvector $\mathbf{p}$ of the $K \times K$ sample covariance matrix, with a corresponding eigenvalue $\mathbf{t}^T\mathbf{t}$:

$$\mathbf{X}^T\mathbf{X}\mathbf{p} = \left(\mathbf{t}^T\mathbf{t}\right)\mathbf{p}$$

The above algorithm computes a single principal component of $\mathbf{X}$ in the form of $\mathbf{t}$ and $\mathbf{p}$. In order to compute a second component, the first component's contributions must be subtracted from $\mathbf{X}$, a step referred to as "deflation":

$$\mathbf{X}' \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T \tag{3.24}$$

Re-application of the NIPALS algorithm to the deflated matrix $\mathbf{X}'$ will then produce the second principal component of the original data matrix $\mathbf{X}$. This process of power iteration and deflation is repeated until an optimal number of components A^* is reached.

PCA for Unsupervised Modeling

From the algebraic properties of PCA, an intuitive geometric picture may be constructed (Figure 3.9). In essence, PCA determines the directions within a data matrix – the principal components – that contain the greatest sources of variation in that matrix (3.9A), where each direction is constrained to be orthogonal to all previous directions. When observations in $\mathbf{X}$ fall into two or more experimental groups, they produce different clustering patterns in the PCA scores $\mathbf{T}$. In cases where the within-group variation in one or more groups is the major source of variation in $\mathbf{X}$, PCA will fail to effectively separate the groups in scores space (3.9B). However, when the between-group variation significantly contributes to the total variation in $\mathbf{X}$, the groups will be well-separated (3.9C). Thus, PCA is a powerful method of identifying major trends among variables in a data matrix, as well as

general relationships between observations, and does not bias its results based on class identity. For a description of methods to quantify separations between experimental groups in PCA scores space, see Chapter 10.

PCA for Outlier Detection

During data handling, it is common practice to exclude outlying observations, when warranted by the analysis, in order to ensure relevant information extraction. In the univariate case, the sample mean $\bar{x}$ and sample standard deviation s may be estimated from a data vector $\mathbf{x}$, which may then be standardized into a set of Mahalanobis distances $\mathbf{d}$:

$$d_n = \frac{x_n - \bar{x}}{s} \quad \forall n \in \mathbb{Z}_1^N \tag{3.25}$$

These distances in $\mathbf{d}$ may be transformed into t statistics and compared to critical values of a t-distribution using a run plot in order to visually detect outliers [14]. In the multivariate case, the sample mean vector $\bar{\mathbf{x}} \in \mathbb{R}^K$ and the sample variance-covariance matrix $\mathbf{S} \in \mathbb{R}^{K \times K}$ of the data matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ must first be computed:

$$\bar{\mathbf{x}} = N^{-1} \sum_{n=1}^{N} \mathbf{x}_n \tag{3.26}$$

$$\mathbf{S} = (N-1)^{-1} \mathbf{X}^T \mathbf{X} \tag{3.27}$$

where $\mathbf{x}_n$ is the n-th observation row vector in $\mathbf{X}$. The set of squared Mahalanobis distances may then be computed as follows [27]:

$$d_n^2 = (\mathbf{x}_n - \bar{\mathbf{x}}) \mathbf{S}^{-1} (\mathbf{x}_n - \bar{\mathbf{x}})^T \quad \forall n \in \mathbb{Z}_1^N \tag{3.28}$$

Once again, each squared distance may be compared to a critical value from a T^2-distribution to assess the probability that its corresponding observation is an outlier. The above procedure fails in practice, because the covariance matrix is severely rank-deficient, and thus non-invertible (i.e. $N \ll K$). To ensure a stable inversion of the covariance matrix, the data matrix may be approximated using PCA (i.e. $\mathbf{X} \approx \mathbf{T}\mathbf{P}^T$), where the number of principal components is less than the rank of the data matrix. Using the orthonormality property of PCA loadings, squared

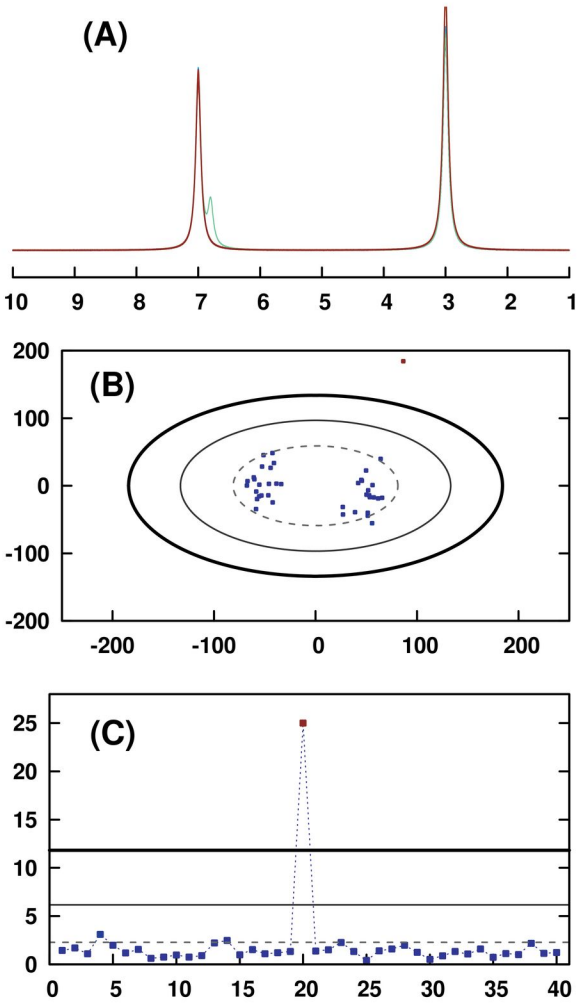

Figure 3.10: PCA for Outlier Testing.
(**A**) Set of 40 spectra, generated identically to those in Figure 3.7, with the exception of observation 20, which contains an extra Lorentzian signal at 6.8 ppm.
(**B**) Principal Component Analysis scores plot of the spectral data, showing the 68.3% (1σ), 95.5% (2σ) and 99.7% (3σ) confidence regions as dashed, thin and bold ellipses, respectively. Point colorings indicate relative squared Mahalanobis distances, ranging from blue to red as d^2 increases.
(**C**) Run plot of squared Mahalanobis distances computed from PCA scores of the spectral data, again illustrating the one-, two- and three-σ thresholds for outlier detection. Point colorings indicate relative squared Mahalanobis distances.

Mahalanobis distances may again be obtained from the PCA scores:

$$d_n^2 = \mathbf{t}_n \left(\frac{\mathbf{T}^T\mathbf{T}}{N-1} \right)^{-1} \mathbf{t}_n^T \quad \forall n \in \mathbb{Z}_1^N \tag{3.29}$$

where $\mathbf{t}_n$ is the n-th row of the scores matrix $\mathbf{T}$. Because the matrix $\mathbf{T}^T\mathbf{T}$ is diagonal, it is trivially inverted and calculation of each d_n^2 is greatly simplified. Mahalanobis distances computed using PCA scores are close approximations of their true values in the original high-dimensional space [27], and provide a means of outlier detection in high-dimensional data (Figure 3.10). Outliers may be visually identified from scatter plots of PCA scores (Figure 3.10B) or from run plots of their corresponding squared Mahalanobis distances (Figure 3.10C). In both cases, the squared Mahalanobis distance is compared to the χ^2 distribution at a preselected significance α and A degrees of freedom [50, 106].

PCA for Multiple Linear Regression

Often, a data matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ is used within the context of a multiple linear regression to determine a set of regression coefficients $\mathbf{B} \in \mathbb{R}^{K \times M}$ that best recapitulate a set of responses $\mathbf{Y} \in \mathbb{R}^{N \times M}$ using the data in $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{3.30}$$

In which case the ordinary least squares (OLS) estimator of the regression coefficients is obtained from inversion of the normal equations [31]:

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.31}$$

The OLS regression coefficients $\hat{\mathbf{B}}$ minimize the sum of squares of the residual matrix, $\|\mathbf{E}\|_F^2$, and are maximum likelihood estimates of $\mathbf{B}$ when the errors are independent and identically normally distributed [31]. However, the fact that $N \ll K$ yet again makes the matrix $\mathbf{X}^T\mathbf{X}$ non-invertible, forcing the analyst down a different path. By replacing the data matrix with its PCA approximation in the regression equation:

$$\mathbf{Y} = \mathbf{TP}^T\mathbf{B} + \mathbf{E}' \tag{3.32}$$

it is possible to obtain OLS estimates of the regression coefficients:

$$\begin{aligned}\hat{\mathbf{B}}' &= (\mathbf{PT}^T\mathbf{TP}^T)^{-1}\mathbf{PT}^T\mathbf{Y} \\ &= \mathbf{P}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y} &(3.33)\\ &= \mathbf{P}\hat{\mathbf{G}} &(3.34)\end{aligned}$$

where the matrix $\hat{\mathbf{G}} \in \mathbb{R}^{A \times M}$ is the set of OLS regression coefficient estimates in PCA scores space:

$$\mathbf{Y} = \mathbf{TG} + \mathbf{E}' \tag{3.35}$$

By computing the least-squares estimates of the regression coefficients in the low-dimensional PCA scores space and projecting those estimates into the original high-dimensional space, this technique of principal component regression (PCR, [51]) sidesteps the curse of dimensionality during estimation. Furthermore, the variances of PCR-estimated coefficients $\hat{\mathbf{B}}'$ are lower than those of the original OLS estimates $\hat{\mathbf{B}}$. The PCR method will fail to obtain useful estimates, however, when response-correlated information in the data matrix is not a major source of variation (cf. Figure 3.9B). In

such cases, methods such as PLS must be used instead.

3.5.2 Partial Least Squares

Partial least squares (PLS) approaches the high-dimensional multiple linear regression problem introduced in equation 3.30 by approximating both $\mathbf{X}$ *and* $\mathbf{Y}$ using bilinear factorizations [104]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.36}$$

$$\begin{aligned}\mathbf{Y} &= \mathbf{U}\mathbf{C}^T + \mathbf{G} \\ &= \mathbf{T}\mathbf{C}^T + \mathbf{F}\end{aligned} \tag{3.37}$$

where the last equality indicates that the $\mathbf{X}$-scores are highly correlated with the $\mathbf{Y}$-scores, and are therefore good predictors of $\mathbf{Y}$. The most commonly applied algorithm for PLS is once again NIPALS-based, and is shown below:

Algorithm 3.2 NIPALS Algorithm for PLS

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}$, $\mathbf{Y} \in \mathbb{R}^{N \times M}$
Output: $\mathbf{t} \in \mathbb{R}^N$, $\mathbf{p} \in \mathbb{R}^K$, $\mathbf{w} \in \mathbb{R}^K$, $\mathbf{u} \in \mathbb{R}^N$, $\mathbf{c} \in \mathbb{R}^M$
1: $\mathbf{u} \sim U_{N \times 1}$ { $\mathbf{u}$ may also be initialized to a column of $\mathbf{Y}$ }
2: **repeat**
3: $\quad \mathbf{w} \propto \mathbf{X}^T\mathbf{u}$
4: $\quad \mathbf{t} \leftarrow \mathbf{X}\mathbf{w}$
5: $\quad \mathbf{c} \leftarrow \frac{\mathbf{Y}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$
6: $\quad \mathbf{u} \leftarrow \frac{\mathbf{Y}\mathbf{c}}{\mathbf{c}^T\mathbf{c}}$
7: **until** $\tau < \varepsilon$
8: $\mathbf{p} \leftarrow \frac{\mathbf{X}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$

where τ equals the score convergence value:

$$\tau = \frac{\|\mathbf{t}^{(k)} - \mathbf{t}^{(k-1)}\|_2}{\|\mathbf{t}^{(k-1)}\|_2} \tag{3.38}$$

and iteration superscripts have been dropped for readability. Backtracking the iteration assignments now produces a different iteration equation:

$$\mathbf{w}^{(k)} \leftarrow \frac{\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}^{(k-1)}}{\|\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}^{(k-1)}\|_2} \tag{3.39}$$

which is the equation for power iteration on the cross-covariance matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. The NIPALS PLS algorithm computes the dominant eigenvector $\mathbf{w}$ of the cross-covariances between the data and

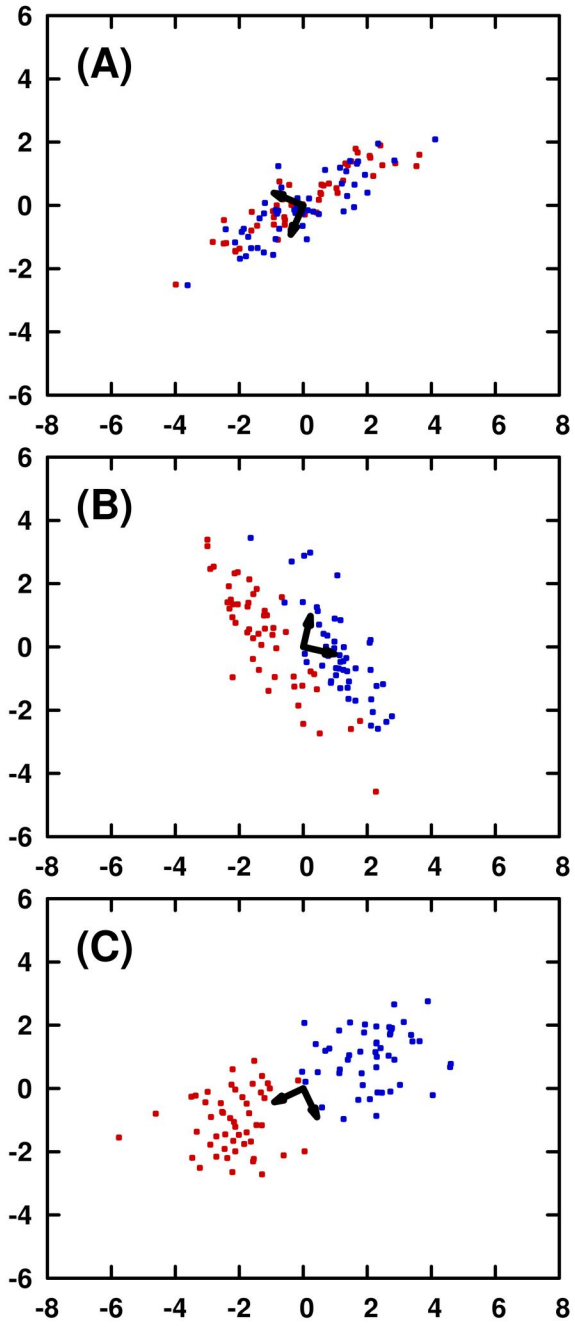

Figure 3.11: PLS Components of Synthetic Bivariate Data.
(**A**) Set of 100 points drawn from a bivariate normal distribution, and the corresponding partial least squares weights computed by eigendecomposition of the points' cross-covariances with the response vector.
(**B**) Set of 100 points drawn from two bivariate normal distributions having different means, where the major source of variation in the data is orthogonal to the direction that separates the two groups. Note the mixing of predictive information between both PLS components.
(**C**) Set of 100 points drawn from two bivariate normal distributions having different means, where the major source of variation is parallel to the direction that separates the groups.

responses (cf. Figure 3.11). As in PCA, computation of subsequent PLS components is achieved by deflating the data and response matrices:

$$\mathbf{X}' \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T \tag{3.40}$$

$$\mathbf{Y}' \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{c}^T \tag{3.41}$$

and re-applying NIPALS on $\mathbf{X}'$ and $\mathbf{Y}'$. Unlike PCA, the PLS loadings $\mathbf{P}$ are not orthogonal, and the $\mathbf{X}$-scores are instead obtained through a linear transformation by a set of non-orthogonal "weights" $\mathbf{W}^* \in \mathbb{R}^{K \times A}$, like so:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{3.42}$$

This allows PLS to be rewritten into the form of a multiple linear regression model:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{W}^*\mathbf{C}^T + \mathbf{F} \\ &= \mathbf{X}\hat{\mathbf{B}}_{PLS} + \mathbf{F} \end{aligned} \tag{3.43}$$

These weights $\mathbf{W}^*$, which directly relate to $\mathbf{X}$, may be computed from the orthonormal weights $\mathbf{W}$ returned from NIPALS through the following transformation [64]:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \tag{3.44}$$

3.5.3 Orthogonal Projections to Latent Structures

The PLS modeling framework expresses the data and response matrices as a pair of low-rank bilinear factorizations, where the data scores $\mathbf{T}$ hold variation in the data that is $\mathbf{Y}$-predictive, as well as variation that compensates for the $\mathbf{Y}$-uncorrelated portion of the data [44]. As a result, PLS models typically require more components than response matrix columns. In other words, $A^* \geq M$ for any PLS model, where the two are equal if and only if no $\mathbf{Y}$-uncorrelated variation is present in $\mathbf{X}$. This presence of "compensatory correlations" in PLS scores confounds interpretation of scores and loadings produced by such non-parsimonious PLS models.

One potential solution proposed to deal with $\mathbf{Y}$-uncorrelated variation, known as orthogonal signal correction (OSC, [5, 97]), removes variation in the data matrix that is not correlated to the responses

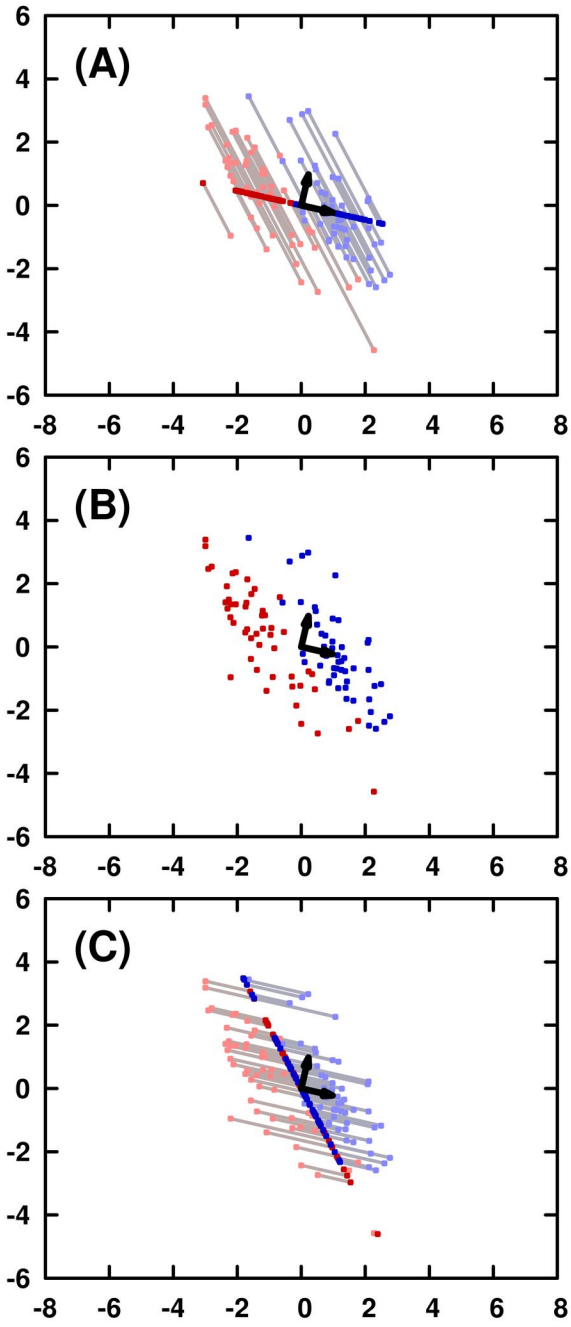

Figure 3.12: Orthogonal Projections of Synthetic Bivariate Data.
(**A**) Projection of the set of points in (**B**) onto their response-predictive (PLS) component.
(**B**) Set of 100 points and their PLS components from Figure 3.11B.
(**C**) Projection of the set of points in (**B**) onto their response-orthogonal (OPLS) component.

using an orthogonal projection that reveals $\mathbf{Y}$-predictive variation:

$$\mathbf{X_p} = \mathbf{X} - \mathbf{T_o}\mathbf{P_o}^T \tag{3.45}$$

$$= \left(\mathbf{I} - \mathbf{T_o}(\mathbf{T_o}^T\mathbf{T_o})^{-1}\mathbf{T_o}^T\right)\mathbf{X} \tag{3.46}$$

where the $\mathbf{Y}$-orthogonal scores $\mathbf{T_o} \in \mathbb{R}^{N \times A_o}$ and loadings $\mathbf{P_o} \in \mathbb{R}^{K \times A_o}$ may be estimated using a variety of algorithms [5]. However, OSC methods tend to suffer from problems of overfitting, and PLS models trained on data matrices that have been filtered by OSC are also at risk of being over-fit [88]. As an alternative to direct methods of orthogonal signal correction, a modified NIPALS PLS algorithm – orthogonal projections to latent structures (OPLS) – was proposed. Instead of removing *all* $\mathbf{Y}$-uncorrelated variation in $\mathbf{X}$ prior to PLS modeling, OPLS only removes $\mathbf{Y}$-uncorrelated variation that interferes with predictive PLS components, effectively partitioning the variation in the data matrix into a set of A_p predictive components and a set of A_o orthogonal components:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{T_o}\mathbf{P_o}^T + \mathbf{E} \tag{3.47}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^T + \mathbf{G} \tag{3.48}$$

$$= \mathbf{T}\mathbf{C}^T + \mathbf{F}$$

The addition of OSC to NIPALS PLS in the form of OPLS is described in the following algorithm:

Algorithm 3.3 NIPALS Algorithm for OPLS

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}$, $\mathbf{Y} \in \mathbb{R}^{N \times M}$
Output: $\mathbf{t} \in \mathbb{R}^{N}$, $\mathbf{p} \in \mathbb{R}^{K}$, $\mathbf{w} \in \mathbb{R}^{K}$, $\mathbf{T_o} \in \mathbb{R}^{N \times a}$, $\mathbf{P_o} \in \mathbb{R}^{K \times a}$, $\mathbf{W_o} \in \mathbb{R}^{K \times a}$, $\mathbf{u} \in \mathbb{R}^{N}$, $\mathbf{c} \in \mathbb{R}^{M}$

1: **for all** $m \in \mathbb{Z}_1^M$ **do**
2: $\quad \mathbf{v}_m \leftarrow \frac{\mathbf{X}^T \mathbf{y}_m}{\mathbf{y}_m^T \mathbf{y}_m}$
3: $\quad \mathbf{V} \leftarrow [\mathbf{V},\ \mathbf{v}_m]$
4: **end for**
5: $\mathbf{u} \sim U_{N \times 1}$ { $\mathbf{u}$ may also be initialized to a column of $\mathbf{Y}$ }
6: done $\leftarrow$ **false**
7: $\mathbf{E} \leftarrow \mathbf{X}$
8: $a \leftarrow 0$
9: **while not** done **do**
10: $\quad$ **repeat**
11: $\quad\quad \mathbf{w} \propto \mathbf{E}^T \mathbf{u}$
12: $\quad\quad \mathbf{t} \leftarrow \mathbf{E}\mathbf{w}$
13: $\quad\quad \mathbf{c} \leftarrow \frac{\mathbf{Y}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}}$
14: $\quad\quad \mathbf{u} \leftarrow \frac{\mathbf{Y}\mathbf{c}}{\mathbf{c}^T \mathbf{c}}$
15: $\quad$ **until** $\tau < \varepsilon$
16: $\quad \mathbf{p} \leftarrow \frac{\mathbf{E}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}}$
17: $\quad \mathbf{z} \leftarrow \mathbf{p}$
18: $\quad \mathbf{z} \leftarrow \mathbf{z} - \frac{\mathbf{v}_m^T \mathbf{z}}{\mathbf{v}_m^T \mathbf{v}_m} \mathbf{v}_m \quad \forall m \in \{1, 2, \ldots M\}$
19: $\quad \mathbf{w_o} \propto \mathbf{z}$
20: $\quad \mathbf{t_o} \leftarrow \mathbf{E}\mathbf{w_o}$
21: $\quad \mathbf{p_o} \leftarrow \frac{\mathbf{E}^T \mathbf{t_o}}{\mathbf{t_o}^T \mathbf{t_o}}$
22: $\quad \lambda \leftarrow \frac{\|\mathbf{z}\|_2}{\|\mathbf{p}\|_2}$
23: $\quad$ **if** $\lambda > \lambda_{th}$ **then**
24: $\quad\quad \mathbf{T_o} \leftarrow [\mathbf{T_o},\ \mathbf{t_o}]$
25: $\quad\quad \mathbf{P_o} \leftarrow [\mathbf{P_o},\ \mathbf{p_o}]$
26: $\quad\quad \mathbf{W_o} \leftarrow [\mathbf{W_o},\ \mathbf{w_o}]$
27: $\quad\quad \mathbf{E} \leftarrow \mathbf{E} - \mathbf{t_o}\mathbf{p_o}^T$
28: $\quad\quad a \leftarrow a + 1$
29: $\quad$ **else**
30: $\quad\quad$ done $\leftarrow$ **true**
31: $\quad$ **end if**
32: **end while**

where τ is again the convergence value for $\mathbf{t}$. The above OPLS algorithm computes one predictive (PLS) component in the vectors $\mathbf{t}$ and $\mathbf{p}$, and a orthogonal components in $\mathbf{T_o}$ and $\mathbf{P_o}$. While the above algorithm appears considerably more complicated than those for PCA or PLS, it is essentially a PLS algorithm (lines 10–16) that has been wrapped in an OSC filter (main "done" loop). At each execution of the main loop, a single PLS component is computed on the current predictive data matrix, $\mathbf{E}$. If a new orthogonal component may be obtained from this PLS component that contains significant variation (i.e. $\lambda > \lambda_{th}$), it is added to the set of orthogonal components and subtracted from $\mathbf{E}$, from which an updated PLS component is computed. As in PCA and PLS, computation of

subsequent OPLS component sets is achieved by first deflating the data and response matrices:

$$\mathbf{X}' \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T - \mathbf{T_o}\mathbf{P_o}^T \tag{3.49}$$

$$\mathbf{Y}' \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{c}^T \tag{3.50}$$

and then re-applying NIPALS on $\mathbf{X}'$ and $\mathbf{Y}'$. Like PLS, the OPLS equations may also be written in the form of a multiple linear regression:

$$\mathbf{Y} = \left(\mathbf{X} - \mathbf{T_o}\mathbf{P_o}^T\right)\mathbf{W}^*\mathbf{C}^T + \mathbf{F} \tag{3.51}$$

It is important to note that OPLS does not outperform PLS in prediction [83], but merely provides a more easily interpretable, parsimonious model for the analyst. In fact, predictions made by OPLS models having A_p predictive components and A_o orthogonal components are identical to those made by PLS models having $A_p + A_o$ components. Also, when the data matrix contains no $\mathbf{Y}$-uncorrelated variation, OPLS and PLS will produce identical models.

O2PLS

While the above NIPALS OPLS algorithm is capable of handling matrix-$\mathbf{Y}$ multivariate regression problems, its creators have championed the O2PLS [89] method instead of OPLS in such situations. Whereas OPLS is a unidirectional (i.e. $\mathbf{X} \mapsto \mathbf{Y}$) regression method, the bidirectional O2PLS method considers neither matrix to be special, and decomposes each matrix into a 'local' or 'unique' component and a 'joint' component (i.e. $\mathbf{X} \leftrightarrow \mathbf{Y}$). The O2PLS modeling method provides an unsupervised means of analyzing relationships between two spectral data matrices, where variation exists in each matrix that is uncorrelated to the other.

3.5.4 Consensus PCA

In analogy to PCA, which seeks directions of maximum variation (loadings) in a data matrix, the consensus PCA (CPCA-W, [99, 77]) seeks a set of "consensus" loadings $\mathbf{P}_b$ that contain a maximum amount of variation in each of the B provided data blocks $\mathbf{X}_b$ while retaining variation that relates each block to the others. In essence, CPCA-W returns a PCA "super-model" for the matrix of

concatenated blocks $\mathbf{X}$ and a PCA block model for each individual block $\mathbf{X}_b$:

$$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_B] \tag{3.52}$$

$$= \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.53}$$

$$\mathbf{X}_b = \mathbf{T}_b {\mathbf{P}_b}^T + \mathbf{E}_b \quad \forall b \in \mathbb{Z}_1^B \tag{3.54}$$

The NIPALS algorithm for CPCA-W, shown below, returns a pair of super-scores $\mathbf{t}$ and super-loadings $\mathbf{p}$ that relate to the matrix of concatenated blocks, as well as a pair of block scores $\mathbf{t}_b$ and block loadings $\mathbf{p}_b$ for each of the B provided blocks:

Algorithm 3.4 NIPALS Algorithm for CPCA-W

Input: $\mathbf{X}_b \in \mathbb{R}^{N \times K_b} \quad \forall b \in \mathbb{Z}_1^B$
Output: $\mathbf{t} \in \mathbb{R}^N, \mathbf{p} \in \mathbb{R}^K, \mathbf{w} \in \mathbb{R}^B, \mathbf{t}_b \in \mathbb{R}^N, \mathbf{p}_b \in \mathbb{R}^K \quad \forall b \in \mathbb{Z}_1^B$

1: $\mathbf{t} \sim U_{N \times 1}$ { $\mathbf{t}$ may also be initialized to a column of $\mathbf{X}$ }
2: **repeat**
3: **for all** $b \in \mathbb{Z}_1^B$ **do**
4: $\mathbf{p}_b \propto {\mathbf{X}_b}^T \mathbf{t}$
5: $\mathbf{t}_b \leftarrow \mathbf{X}_b \mathbf{p}_b$
6: **end for**
7: $\mathbf{R} \leftarrow [\mathbf{t}_1, \ldots \mathbf{t}_B]$
8: $\mathbf{w} \propto \frac{{\mathbf{R}}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}}$
9: $\mathbf{t} \leftarrow \mathbf{R}\mathbf{w}$
10: **until** $\tau < \varepsilon$
11: $\mathbf{p}_b \leftarrow \frac{{\mathbf{X}_b}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}} \quad \forall b \in \mathbb{Z}_1^B$
12: $\mathbf{p}^T \leftarrow [{\mathbf{p}_1}^T, \ldots {\mathbf{p}_B}^T]$

where τ relates to the convergence value of the super-scores $\mathbf{t}$ as in standard PCA. Computation of subsequent components requires the deflation of each data block, like so:

$$\mathbf{X}'_b \leftarrow \mathbf{X}_b - \mathbf{t}{\mathbf{p}_b}^T \quad \forall b \in \mathbb{Z}_1^B \tag{3.55}$$

When block scaling is applied to each data block, the super-scores and super-loadings produced by CPCA-W of the blocks will be equivalent to scores and loadings from PCA of the concatenated matrix [99, 77]. Thus, the super-loadings ($\mathbf{p}$) from CPCA-W are again the set of eigenvectors of the sample covariance matrix of $\mathbf{X}$. The loadings of each individual block are the projections of that block onto the super-scores (Algorithm 3.4, line 11). As a result, each block model describes block-specific variation that is correlated to the consensus directions of the combined set of blocks.

3.5.5 Multiblock PLS

Several extensions of PLS [99] have been proposed for problems of high-dimensional multiple linear regression of multiple data blocks against a single matrix of responses. However, one particular extension, known as multiblock PLS (MB-PLS), was described [96] that has several attractive features. Like CPCA-W, MB-PLS constructs a super-model that relates the concatenated set of B blocks to the response matrix $\mathbf{Y}$, but concomitantly breaks each block into its own model that describes that block's contribution to $\mathbf{Y}$-prediction:

$$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_B] \tag{3.56}$$

$$= \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.57}$$

$$\mathbf{X}_b = \mathbf{T}_b {\mathbf{P}_b}^T + \mathbf{E}_b \quad \forall b \in \mathbb{Z}_1^B \tag{3.58}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^T + \mathbf{G} \tag{3.59}$$

$$= \mathbf{T}\mathbf{C}^T + \mathbf{F}$$

While the practical calculation of MB-PLS models is usually performed by extracting block models from a PLS model trained on $\mathbf{X}$, the original algorithm for MB-PLS is of the NIPALS-type, and is shown below:

Algorithm 3.5 NIPALS Algorithm for MB-PLS

Input: $\mathbf{X}_b \in \mathbb{R}^{N \times K_b} \quad \forall b \in \mathbb{Z}_1^B, \mathbf{Y} \in \mathbb{R}^{N \times M}$
Output: $\mathbf{t} \in \mathbb{R}^N, \mathbf{p} \in \mathbb{R}^K, \mathbf{u} \in \mathbb{R}^N, \mathbf{c} \in \mathbb{R}^M, \mathbf{w} \in \mathbb{R}^B, \mathbf{t}_b \in \mathbb{R}^N, \mathbf{p}_b \in \mathbb{R}^K, \mathbf{w}_b \in \mathbb{R}^K \quad \forall b \in \mathbb{Z}_1^B$

```
u ~ U_{N×1} { u may also be initialized to a column of Y }
repeat
    for all b ∈ Z_1^B do
        w_b ∝ X_b^T u
        t_b ← X_b w_b
    end for
    R ← [t_1, ... t_B]
    w ∝ R^T u
    t ← R w
    c ← (Y^T t) / (t^T t)
    u ← (Y c) / (c^T c)
until τ < ε
p_b ← (X_b^T t) / (t^T t)   ∀b ∈ Z_1^B
p^T ← [p_1^T, ... p_B^T]
```

where τ once again relates to the convergence value of the super-scores $\mathbf{t}$. Computation of subsequent

components requires the deflation of each data block using super-scores [96], like so:

$$\mathbf{X}_b' \leftarrow \mathbf{X}_b - \mathbf{t}\mathbf{p}_b{}^T \quad \forall b \in \mathbb{Z}_1^B \tag{3.60}$$

$$\mathbf{Y}' \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{c}^T \tag{3.61}$$

It can be shown [99] that MB-PLS using block scaling and super-score deflation produces an identical super-model to that obtained by PLS modeling of the concatenated matrix $\mathbf{X}$, which indicates that the super-weights produced by MB-PLS are still eigenvectors of the matrix of cross-covariances between $\mathbf{X}$ and $\mathbf{Y}$. Thus, MB-PLS provides the analyst with a standard PLS model that describes how the joint data in $\mathbf{X}$ predict $\mathbf{Y}$, as well as how each individual data block $\mathbf{X}_b$ contributes to the joint prediction.

3.5.6 Multiblock OPLS

While MB-PLS provides a powerful framework for multiblock multivariate regression problems, it suffers from the same shortcomings of PLS when $\mathbf{Y}$-uncorrelated variation exists in one or more data blocks [61, 60]. To address this flaw in MB-PLS, the generalized OnPLS multiblock modeling framework, which extends O2PLS to B data blocks, was developed [60]. Like in O2PLS, no matrix in OnPLS is special, and all matrices are regressed against all others in a complete association graph (Figure 3.13A). While such complete connectivity may be useful during unsupervised modeling, it is an over-complication in the multiblock regression schemes normally handled in metabolomics (cf. Chapter 9). Because each data block must only predict a single block of responses (Figure 3.11B), the recently developed multiblock OPLS (MB-OPLS) method is a more suitable candidate.

Multiblock OPLS decomposes each block $\mathbf{X}_b$ into a set of $\mathbf{Y}$-predictive scores and loadings, as well as a set of $\mathbf{Y}$-uncorrelated scores and loadings that would normally interfere with MB-PLS

predictions:

$$\mathbf{X} = [\mathbf{X}_1, \ldots \mathbf{X}_B] \tag{3.62}$$

$$= \mathbf{T}\mathbf{P}^T + \mathbf{T_o}\mathbf{P_o}^T + \mathbf{E} \tag{3.63}$$

$$\mathbf{X}_b = \mathbf{T}_b\mathbf{P}_b{}^T + \mathbf{T}_{\mathbf{o}b}\mathbf{P}_{\mathbf{o}b}{}^T + \mathbf{E}_b \quad \forall b \in \mathbb{Z}_1^B \tag{3.64}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^T + \mathbf{G} \tag{3.65}$$

$$= \mathbf{T}\mathbf{C}^T + \mathbf{F}$$

The MB-OPLS algorithm, described in greater detail in Chapter 9, is based on NIPALS MB-PLS with an added OPLS-type OSC filter that removes $\mathbf{Y}$-uncorrelated variation from super-loadings. Deflation of predictive and orthogonal components from each block is achieved using super-scores, as in the above MB-PLS algorithm. The complete matrix-$\mathbf{Y}$ MB-OPLS algorithm follows on the next page.

Algorithm 3.6 NIPALS Algorithm for MB-OPLS

Input: $\mathbf{X}_b \in \mathbb{R}^{N \times K_b} \quad \forall b \in \mathbb{Z}_1^B, \mathbf{Y} \in \mathbb{R}^{N \times M}$
Output: $\mathbf{t} \in \mathbb{R}^N, \mathbf{p} \in \mathbb{R}^K, \mathbf{u} \in \mathbb{R}^N, \mathbf{c} \in \mathbb{R}^M, \mathbf{T_o} \in \mathbb{R}^{N \times a}, \mathbf{P_o} \in \mathbb{R}^{K \times a}, \mathbf{w} \in \mathbb{R}^B,$
$\mathbf{t}_b \in \mathbb{R}^N, \mathbf{p}_b \in \mathbb{R}^K, \mathbf{w}_b \in \mathbb{R}^K, \mathbf{T_o}_b \in \mathbb{R}^{N \times a}, \mathbf{P_o}_b \in \mathbb{R}^{K \times a} \quad \forall b \in \mathbb{Z}_1^B$

$\mathbf{u} \sim U_{N \times 1}$ { $\mathbf{u}$ may also be initialized to a column of $\mathbf{Y}$ }
done $\leftarrow$ **false**
$a \leftarrow 0$
$\mathbf{E}_b \leftarrow \mathbf{X}_b \quad \forall b \in \mathbb{Z}_1^B$
while not done **do**
  repeat
    for all $b \in \mathbb{Z}_1^B$ **do**
      $\mathbf{w}_b \propto {\mathbf{E}_b}^T \mathbf{u}$
      $\mathbf{t}_b \leftarrow \mathbf{E}_b \mathbf{w}_b$
    end for
    $\mathbf{R} \leftarrow [\mathbf{t}_1, \dots \mathbf{t}_B]$
    $\mathbf{w} \propto \mathbf{R}^T \mathbf{u}$
    $\mathbf{t} \leftarrow \mathbf{R}\mathbf{w}$
    $\mathbf{c} \leftarrow \frac{\mathbf{Y}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}}$
    $\mathbf{u} \leftarrow \frac{\mathbf{Y}\mathbf{c}}{\mathbf{c}^T \mathbf{c}}$
  until $\tau \leq \varepsilon$
  $\mathbf{p}_b \leftarrow \frac{{\mathbf{E}_b}^T \mathbf{t}}{\mathbf{t}^T \mathbf{t}} \quad \forall b \in \mathbb{Z}_1^B$
  $\mathbf{p}^T \leftarrow [{\mathbf{p}_1}^T, \dots {\mathbf{p}_B}^T]$
  $\mathbf{z} \leftarrow \mathbf{p}$
  $\mathbf{z} \leftarrow \mathbf{z} - \frac{\mathbf{v}_m^T \mathbf{z}}{\mathbf{v}_m^T \mathbf{v}_m} \mathbf{v}_m \quad \forall m \in \mathbb{Z}_1^M$
  $\mathbf{w_o} \propto \mathbf{z}$
  $\mathbf{t_o} \leftarrow [\mathbf{E}_1, \dots \mathbf{E}_B] \mathbf{w_o}$
  $\mathbf{p_o}_b \leftarrow \frac{{\mathbf{E}_b}^T \mathbf{t_o}}{{\mathbf{t_o}}^T \mathbf{t_o}} \quad \forall b \in \mathbb{Z}_1^B$
  $\mathbf{t_o}_b \leftarrow \frac{\mathbf{E}_b \mathbf{p_o}_b}{{\mathbf{p_o}_b}^T \mathbf{p_o}_b} \quad \forall b \in \mathbb{Z}_1^B$
  ${\mathbf{p_o}}^T \leftarrow [{\mathbf{p_o}_1}^T, \dots {\mathbf{p_o}_B}^T]$
  $\lambda \leftarrow \frac{\|\mathbf{z}\|_2}{\|\mathbf{p}\|_2}$
  if $\lambda < \lambda_{th}$ **then**
    $\mathbf{T_o} \leftarrow [\mathbf{T_o},\ \mathbf{t_o}]$
    $\mathbf{P_o} \leftarrow [\mathbf{P_o},\ \mathbf{p_o}]$
    $\mathbf{W_o} \leftarrow [\mathbf{W_o},\ \mathbf{w_o}]$
    $\mathbf{T_o}_b \leftarrow [\mathbf{T_o}_b,\ \mathbf{t_o}_b] \quad \forall b \in \mathbb{Z}_1^B$
    $\mathbf{P_o}_b \leftarrow [\mathbf{P_o}_b,\ \mathbf{p_o}_b] \quad \forall b \in \mathbb{Z}_1^B$
    $\mathbf{E}_b \leftarrow \mathbf{E}_b - \mathbf{t_o}{\mathbf{p_o}_b}^T \quad \forall b \in \mathbb{Z}_1^B$
    $a \leftarrow a + 1$
  else
    done $\leftarrow$ **true**
  end if
end while

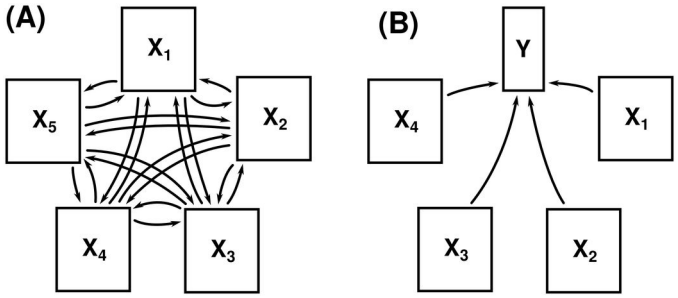

Figure 3.13: Association Graphs for OnPLS and MB-OPLS. (**A**) Completely connected graph of nPLS and OnPLS regression models, in which no single matrix is considered unique. (**B**) Sparsely connected acyclic graph of MB-PLS and MB-OPLS regression models, where each data block $\mathbf{X}_b$ predicts the unique matrix $\mathbf{Y}$.

Computation of another predictive component requires the deflation of each data block by both the predictive super-scores $\mathbf{t}$ and the orthogonal super-scores $\mathbf{T_o}$, similar to the super-score deflation method in MB-PLS [96]:

$$\mathbf{X}'_b \leftarrow \mathbf{X}_b - \mathbf{t}{\mathbf{p}_b}^T - \mathbf{T_o}{\mathbf{P_o}_b}^T \quad \forall b \in \mathbb{Z}_1^B \tag{3.66}$$

$$\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{c}^T \tag{3.67}$$

In the same manner that block factors of CPCA-W and MB-PLS may be obtained from PCA and PLS models – respectively – of the concatenated matrix of blocks [99, 77], MB-OPLS block scores and loadings may be obtained from an OPLS model trained on the concatenated matrix $\mathbf{X}$. The following algorithm describes the steps required to extract block-level models from super-model scores and loadings.

Algorithm 3.7 Extraction of MB-OPLS Factors from OPLS

Input: $\mathbf{X}_b \in \mathbb{R}^{N \times K_b} \quad \forall b \in \mathbb{Z}_1^B, \mathbf{Y} \in \mathbb{R}^{N \times M}$
Output: $\mathbf{t} \in \mathbb{R}^N, \mathbf{p} \in \mathbb{R}^K, \mathbf{u} \in \mathbb{R}^N, \mathbf{c} \in \mathbb{R}^M, \mathbf{T_o} \in \mathbb{R}^{N \times a}, \mathbf{P_o} \in \mathbb{R}^{K \times a}$,
$\mathbf{t}_b \in \mathbb{R}^N, \mathbf{p}_b \in \mathbb{R}^K, \mathbf{T_o}_b \in \mathbb{R}^{N \times a}, \mathbf{P_o}_b \in \mathbb{R}^{K \times a} \quad \forall b \in \mathbb{Z}_1^B$

$\{\mathbf{t}, \mathbf{p}, \mathbf{u}, \mathbf{c}, \mathbf{T_o}, \mathbf{P_o}\} \leftarrow \text{OPLS}(\mathbf{X}, \mathbf{Y})$
for all $b \in \mathbb{Z}_1^B$ **do**
  for all $a_o \in \mathbb{Z}_1^a$ **do**
    $\mathbf{t_o} \leftarrow \mathbf{T_o}_{a_o}$ { Extract the a_o-th column of $\mathbf{T_o}$ }
    $\mathbf{w_o}_b \leftarrow [\mathbf{W_o}_{a_o}]_b$ { Extract the b-th block of of $\mathbf{W_o}_{a_o}$ }
    $\mathbf{t_o}_b \leftarrow \mathbf{X}_b\mathbf{w_o}_b$
    $\mathbf{p_o}_b \leftarrow \frac{{\mathbf{X}_b}^T\mathbf{t_o}}{{\mathbf{t_o}}^T\mathbf{t_o}}$
    $\mathbf{P_o}_b \leftarrow [\mathbf{P_o}_b, \mathbf{p_o}_b]$
    $\mathbf{T_o}_b \leftarrow [\mathbf{T_o}_b, \mathbf{t_o}_b]$
    $\mathbf{X}_b \leftarrow \mathbf{X}_b - \mathbf{t_o}{\mathbf{p_o}_b}^T$
  end for
  $\mathbf{w}_b \propto {\mathbf{X}_b}^T\mathbf{u}$
  $\mathbf{t}_b \leftarrow \mathbf{X}_b\mathbf{w}_b$
  $\mathbf{p}_b \leftarrow \frac{{\mathbf{X}_b}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$
  $\mathbf{X}_b \leftarrow \mathbf{X}_b - \mathbf{t}{\mathbf{p}_b}^T$
end for

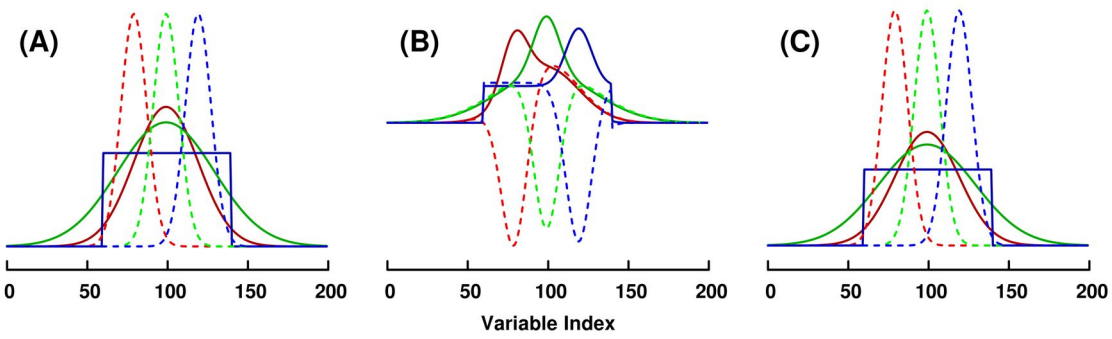

Figure 3.14: Comparison of MB-PLS and MB-OPLS Loadings.
(**A**) Original predictive (solid lines) and orthogonal (dashed lines) block loadings used to construct a toy three-block data structure having 100 observations and 200 variables per block. (**B**) Resulting MB-PLS block loadings from a two-component model, illustrating how the PLS algorithm mixes predictive information between multiple components in the presence of orthogonal variation. (**C**) Resulting MB-OPLS predictive (solid) and orthogonal (dashed) block loadings from a one-component $(1+1)$ model, effectively illustrating how MB-OPLS achieves the same segregation of predictive and orthogonal variation as OPLS and OnPLS on a per-block basis.

For each of the B data blocks, the extraction procedure first computes orthogonal block loadings and scores, and deflates them from the block. Then, predictive block scores and loadings are computed using the method outlined previously [99] for extracting MB-PLS block components from a PLS model (steps 12–14). The resulting block scores and loadings from algorithm 3.7 are identical to those produced by algorithm 3.6.

3.6 Validation

Application of the above bilinear factorization methods to one or more spectral data matrices yields valuable insights into both general chemical trends and relationships (e.g. from PCA) and response-predictive spectral features (e.g. from OPLS) in those matrices. However, wanton use of these multivariate methods without validation or knowledge of algorithmic intent can quickly lead to statistically insignificant conclusions about the underlying chemistry. The NIPALS-based algorithms described above are highly numerically stable, even in the presence of multicollinearity, noise, and missing data [104, 2]. This numerical stability almost guarantees that PCA, PLS and OPLS will return a set of scores and loadings, even when those scores and loadings are only based on a small fraction of the total variation in the data. PLS and OPLS return biased regression coefficient estimates ($\hat{\mathbf{B}}_{PLS}$) and force separation based on responses in scores space. OPLS is especially adept at forcing scores-space separation, because its integrated OSC filter removes systematic data matrix variation that does not "agree" with the responses. These powerful modeling features make PLS and

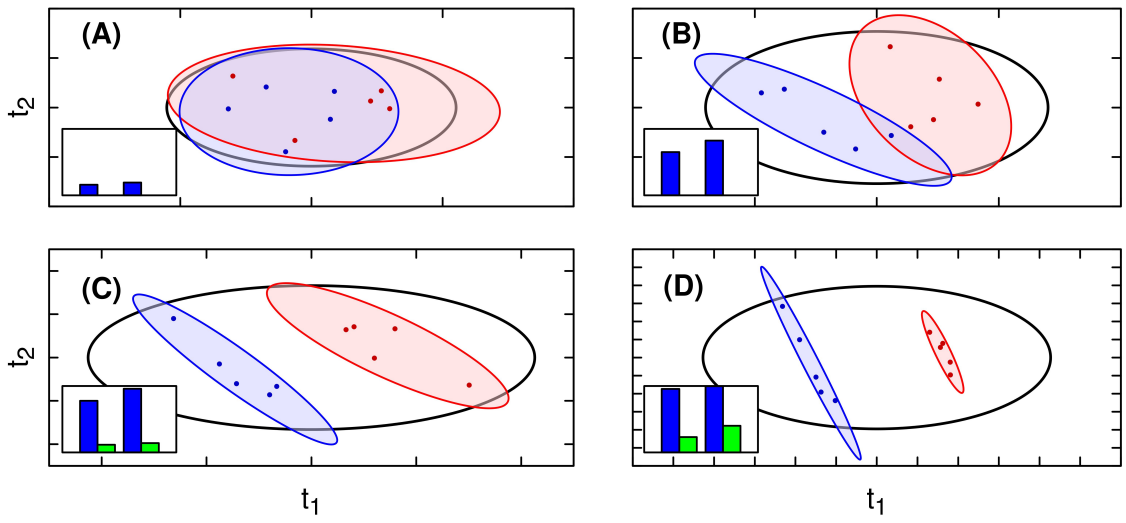

Figure 3.15: Demonstration of PLS Overfit Based on Variable Count.
(**A**) Scores from a two-component PLS model of unit-variance scaled Gaussian white noise ($N = 10$, $K = 5$). Inset: R^2_Y (blue) and Q^2 (green) statistics for each component in the model, computed from 100 iterations of seven-fold Monte Carlo cross-validation. (**B**) Same scenario as (**A**) with $K = 10$. (**C**) Same scenario as (**A**) with $K = 20$. (**D**) Same scenario as (**A**) with $K = 100$.

OPLS fully capable of producing results based on noise alone, if so requested [98]. As the number of variables in the data increases over the number of observations, the danger of overfitting also increases (Figure 3.15). In effect, the probability of observing correlations to the responses increases with variable count, just as the probability of observing long runs of heads or tails increases with the number of fair coin tosses.

In chemometric studies of spectroscopic datasets, where $N \ll K$, the tendency of bilinear models to over-fit must be balanced by rigorous application of several validation methods. When sufficient validation is lacking, any conclusions drawn from these models should automatically be treated as suspect from a statistical viewpoint. Therefore, all efforts must be taken during experimental design, data collection and handling in order to obtain data and models that acceptably withstand cross-validation.

3.6.1 Explained Variation

In general, the amount of variation explained by a bilinear factorization of a matrix is quantified by its sum of squares:

$$SS\left\{\mathbf{tp}^T\right\} = \left\|\mathbf{tp}^T\right\|_F^2$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm, which will be used as a shorthand for the sum of squares. Because the raw sum of squares is not scale-invariant, it is normally reported as a ratio relative to the total sum of squares. In the context of a PCA (equation 3.20), this ratio is referred to as R_X^2 (or simply R^2):

$$R_X^2 = \frac{SS_{\text{fit}}}{SS_{\text{total}}} = \frac{\left\|\mathbf{T}\mathbf{P}^T\right\|_F^2}{\|\mathbf{X}\|_F^2} \tag{3.68}$$

$$= 1 - \frac{SS_{\text{err}}}{SS_{\text{total}}} = 1 - \frac{\|\mathbf{E}\|_F^2}{\|\mathbf{X}\|_F^2} \tag{3.69}$$

where the second set of equalities arises from the fact that PCA loadings are eigenvectors of $\mathbf{X}^T\mathbf{X}$ and right singular vectors of $\mathbf{X}$ [51]. The cumulative R_X^2 statistic may be broken into sums of per-component R_X^2 statistics, due to the orthonormality properties of principal components:

$$R_X^2 = \sum_{a=1}^{A} \frac{\left\|\mathbf{t}_a {\mathbf{p}_a}^T\right\|_F^2}{\|\mathbf{X}\|_F^2} \tag{3.70}$$

In the case of PLS modeling (equation 3.37), a second ratio exists (R_Y^2) that quantifies the amount of variation explained in the responses:

$$R_Y^2 = \frac{\left\|\mathbf{T}\mathbf{C}^T\right\|_F^2}{\|\mathbf{Y}\|_F^2} \tag{3.71}$$

$$= 1 - \frac{\|\mathbf{F}\|_F^2}{\|\mathbf{Y}\|_F^2} \tag{3.72}$$

Similar expressions exist for the predictive and orthogonal factorizations of $\mathbf{X}$ in OPLS models, which are referred to as $R_{X,p}^2$ and $R_{X,o}^2$, respectively. While these R^2 statistics provide valuable first insights into the amount of variation that may be explained by a multivariate model, they are gross over-estimations of model reliability and should not be used during model selection as a means of determining the optimal component count, A^*. Cross-validatory methods discussed in later sections provide more effective means of identifying A^* during model training.

3.6.2 External Cross-validation

In its most traditional form, cross-validation involves the division of N observations in a dataset (i.e. $\mathbf{X}$ and $\mathbf{Y}$) into a training set ($\mathbf{X}_t$ and $\mathbf{Y}_t$) having N_t observations and a validation set ($\mathbf{X}_v$ and $\mathbf{Y}_v$) having N_v observations. The training and validation datasets are then processed and treated

separately using identical methods, and models are constructed on the training dataset. In the case of PLS modeling, the analyst will arrive at the following equation:

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{W}^* \mathbf{C}^T + \mathbf{F} \tag{3.73}$$

Assessment of model reliability is then performed by estimating the responses of the validation dataset using the trained model, like so:

$$\hat{\mathbf{Y}}_v = \mathbf{X}_v \mathbf{W}^* \mathbf{C}^T \tag{3.74}$$

where the predicted residual sum of squares (PRESS) statistic is now readily computable from the sum of squares of the difference between true and estimated validation-set responses:

$$\text{PRESS} = \left\| \mathbf{Y}_v - \hat{\mathbf{Y}}_v \right\|_F^2 \tag{3.75}$$

Like any other sum of squares measure, PRESS depends on the magnitudes of the values in $\mathbf{Y}$. Thus, a scale-invariant reporter of reliability (Q^2) is obtained from the PRESS statistic as follows:

$$Q^2 = 1 - \frac{\text{PRESS}}{SS_{\text{total}}} = 1 - \frac{\left\| \mathbf{Y}_v - \hat{\mathbf{Y}}_v \right\|_F^2}{\left\| \mathbf{Y}_v \right\|_F^2} \tag{3.76}$$

This model reliability statistic is often referred to as a "cross-validated" R^2 statistic, and provides a relative measure of how well a given model will generalize to the estimation of future observations. Like R^2 statistics, Q^2 is also computable on a per-component basis.

When the values in $\mathbf{Y}$ do not vary continuously, but instead hold binary class membership information, it is possible for Q^2 – as defined by the previous equation – to under-estimate model reliability [100]. In the case of two-class discrimination, Q^2 quadratically penalizes values in $\hat{\mathbf{y}}$ that are beyond the class labels in $\mathbf{y}$. In more concrete terms, a cross-validation predicted class label of 1.5 for a true class label of 1.0 should incur no penalty, as it represents an unambiguous prediction. However, the quadratic nature of Q^2 penalizes such results. When PLS and OPLS are used for binary class discrimination, the more suitable "discriminant Q^2" (DQ^2) is a more suitable metric of reliability:

$$DQ^2 = 1 - \frac{\text{PRESSD}}{SS_{\text{total}}} \tag{3.77}$$

where PRESSD represents the discriminant PRESS statistic:

$$\text{PRESSD} = \sum_{n=1}^{N} \begin{cases} 0 & \text{if } y_n = 1 \text{ and } \hat{y}_n > 1 \\ 0 & \text{if } y_n = 0 \text{ and } \hat{y}_n < 0 \\ (y_n - \hat{y}_n)^2 & \text{else} \end{cases} \tag{3.78}$$

In effect, PRESSD computes the sum of all squared $\mathbf{y}$-residuals that represent a potentially ambiguous classification [100]. The same result may be achieved by appropriately bounding the values within $\hat{\mathbf{y}}$ to the class label extents in $\mathbf{y}$, followed by the use of standard PRESS and Q^2 calculations. This alternative strategy has the added benefit of ensuring that the prediction sum of squares $\|\hat{\mathbf{y}}\|_F^2$ never exceeds the total sum of squares $\|\mathbf{y}\|_F^2$, which is a requirement for CV-ANOVA, discussed below.

3.6.3 Internal Cross-validation

Supervised Models

Due to the severely limited number of observations in most chemometric studies, the practice of external cross-validation is rare, and all observations are usually retained for model training. Nevertheless, model reliability statistics may be obtained from the similar practice of *internal* cross-validation. In any internal cross-validation scheme, the N observations of a given dataset are divided into G groups, referred to as G-fold or leave-n-out cross-validation (where $n = \lfloor N/G \rfloor$). Each group is then left out in turn, and its responses are estimated from a model trained on the remainder of the data. To more succinctly introduce internal cross-validation procedures, the set of all partitionings of N elements into G groups, denoted $\mathbb{P}(N, G)$, shall be introduced:

$$\mathbb{P}(N, G) \equiv \left\{ \mathbf{p} \mid \mathbf{p} \in \mathbb{Z}^N \wedge 1 \leq p_n \leq G \quad \forall n \in \mathbb{Z}_1^N \right\} \tag{3.79}$$

As an example, one member of the set $\mathbb{P}(7, 3)$ is $(1, 2, 3, 1, 2, 3, 1)$.[2] Given a partitioning $\boldsymbol{\sigma} \in \mathbb{P}(N, G)$, the following algorithm demonstrates the computation of Q^2 for a single PLS component:

[2] This type of partitioning actually has a name, and is often seen in practical cross-validation schemes. It is colloquially referred to as a "Venetian blinds" partitioning of seven observations into three groups.

Algorithm 3.8 Internal PLS Component Cross-validation

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}, \mathbf{Y} \in \mathbb{R}^{N \times M}, G,\ \boldsymbol{\sigma} \in \mathbb{P}(N, G)$
Output: $Q^2 \in \mathbb{R}$
1: **for all** $g \in \mathbb{Z}_1^G$ **do**
2: $\quad n_g \leftarrow \{n \mid n \in \mathbb{Z}_1^N \wedge \sigma_n = g\}$
3: $\quad \{\mathbf{t}, \mathbf{p}, \mathbf{u}, \mathbf{c}, \mathbf{w}\} \leftarrow \text{PLS}(\mathbf{X}^{(-n_g)}, \mathbf{Y}^{(-n_g)})$ { $\mathbf{X}^{(-n_g)}$, $\mathbf{Y}^{(-n_g)}$ contain no rows from group g }
4: $\quad \mathbf{B}^{(n_g)} \leftarrow \frac{\mathbf{w}\mathbf{c}^T}{\mathbf{p}^T\mathbf{w}}$
5: **end for**
6: **for all** $n \in \mathbb{Z}_1^N$ **do**
7: $\quad \hat{\mathbf{y}}_n \leftarrow \mathbf{x}_n \mathbf{B}^{(\sigma_n)}$ { $\hat{\mathbf{y}}_n$, $\mathbf{x}_n$ are the n-th rows of $\hat{\mathbf{Y}}$, $\mathbf{X}$ }
8: **end for**
9: $Q^2 \leftarrow 1 - \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2}{\|\mathbf{Y}\|_F^2}$

In the simplest case where $G = N$, known as leave-one-out cross-validation (LOOCV), only one observation is left out at a time. It has been shown, however, that leave-one-out methods do not identify optimal models as effectively as leave-n-out methods as N increases [74], so G should be less than the number of observations whenever possible. Within a leave-n-out cross-validation of N observations, there are $\binom{N}{n}$ different ways to partition the dataset into the desired number of groups. A complete cross-validation would require the evaluation of all possible partitions, which is computationally intractable even for small (e.g. $N \geq 20$) datasets. While it is possible to arbitrarily select a single partitioning, such as a regular pattern of group assignment, it is much more attractive to randomly resample a number of partitionings (n_p) from the set of $\binom{N}{n}$ possibilities ($\boldsymbol{\sigma} \sim \mathbb{P}(N, G)$) in a Monte Carlo leave-n-out cross-validation approach [110]. Monte Carlo cross-validation offers the possibility of assigning confidence regions to reported Q^2 values for a given dataset, which provides the analyst with further information on model reliability estimates.

In practice, per-component Q^2 statistics provide a means of determining A^*, the optimal component count. When new components are added to the model that fail to reliably estimate the responses under cross-validation, their Q^2 values will become negative. Thus, a practical rule during model training is to only retain components having positive Q^2 statistics.

Unsupervised Models

Internal cross-validation of unsupervised PCA models poses a unique challenge in comparison to PLS and OPLS, as it does not involve the prediction of a set of responses [33, 56, 37]. Eshghi [37] provides a more comprehensive review of internal cross-validation practices for PCA models. In short, leave-n-out cross-validation of PCA models requires that both observations *and* variables

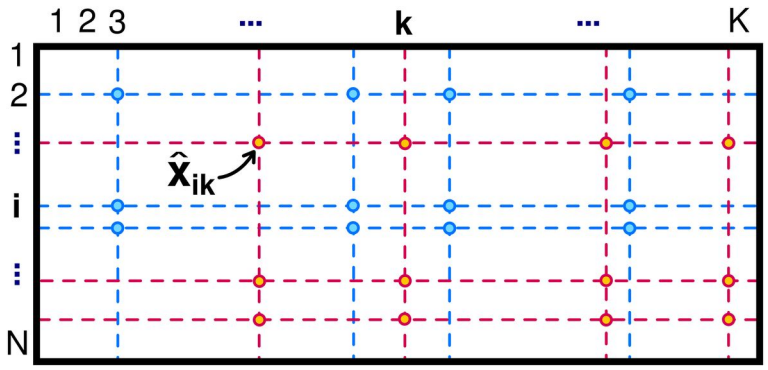

Figure 3.16: Partitioning in Leave-n-out PCA Cross-validation.
Graphical illustration of two different partitionings (red and blue) of a data matrix. Each partitioning (group) requires the computation of a set of scores ($\hat{\mathbf{t}}$) and loadings ($\hat{\mathbf{p}}$) in order to estimate it's left-out values, indicated by filled circles. Estimation of $\hat{x}_{ik}$ requires computation of $\hat{\mathbf{t}}$ with variable k left out and $\hat{\mathbf{p}}$ with observation i left out. The value of $\hat{x}_{ik}$ is then obtained from the (i,k)-th index of $\hat{\mathbf{t}}\hat{\mathbf{p}}^T$

be partitioned into G groups. The resulting pair of partitionings allows groups of data matrix elements to be left out and recomputed during cross-validation. For each group, a score vector $\hat{\mathbf{t}}$ is computed after leaving out variables in the group, and a loading vector $\hat{\mathbf{p}}$ is computed after leaving out observations in the group. The outer product of $\hat{\mathbf{t}}$ and $\hat{\mathbf{p}}$ is then used in estimating the data matrix elements located at the intersections of the left-out variables and observations (Figure 3.16). The following algorithm demonstrates the computation of Q^2 for a single principal component, given a row partitioning $\boldsymbol{\sigma}$ and a column partitioning $\boldsymbol{\rho}$:

Algorithm 3.9 Internal PCA Component Cross-validation

Input: $\mathbf{X} \in \mathbb{R}^{N \times K},\ G,\ \boldsymbol{\sigma} \in \mathbb{P}(N,G), \boldsymbol{\rho} \in \mathbb{P}(K,G)$
Output: $Q^2 \in \mathbb{R}$

for all $g_1 \in \mathbb{Z}_1^G$ **do**
  $n_g \leftarrow \{n \mid n \in \mathbb{Z}_1^N \wedge \sigma_n = g_1\}$
  $\hat{\mathbf{p}} \leftarrow \text{PCA}(\mathbf{X}^{(-n_g)})$ { $\mathbf{X}^{(-n_g)}$ contains no rows from group g}
  for all $g_2 \in \mathbb{Z}_1^G$ **do**
    $k_g \leftarrow \{k \mid k \in \mathbb{Z}_1^K \wedge \rho_k = g_2\}$
    $\hat{\mathbf{t}} \leftarrow \text{PCA}(\mathbf{X}^{(-k_g)})$ { $\mathbf{X}^{(-k_g)}$ contains no columns from group g}
    $\hat{\mathbf{t}} \leftarrow \hat{\mathbf{t}}\sqrt{\frac{K}{K-|k_g|}}$
    for all $n \in n_g$ **do**
      for all $k \in k_g$ **do**
        $\hat{\mathbf{X}}_{n,k} \leftarrow \hat{\mathbf{t}}_n \hat{\mathbf{p}}_k$
      end for
    end for
  end for
end for
$Q^2 \leftarrow 1 - \frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_F^2}{\|\mathbf{X}\|_F^2}$

3.6.4 Response Permutation Testing

While internal cross-validation metrics such as Q^2, number of mis-classifications, and AUROC [98] are a good first insight into model reliability, they do not provide a robust mechanism for discriminating between low- and high-quality models. Methods that report a degree of statistical significance in the form of a p value are preferred, as they may be compared to a threshold (e.g. $\alpha = 0.05$). One such method, known as response permutation testing, establishes a distribution of models representing the null hypothesis (H_0) that no relationship exists between the data and responses [98]. For a number of iterations, the rows of the response matrix $\mathbf{Y}$ are randomly permuted to yield a null response matrix $\tilde{\mathbf{Y}}$, upon which a supervised PLS or OPLS model is trained. The set of models generated after response permutation – more specifically their R^2 and Q^2 parameters – may be compared against those of the original model to obtain a p value. Provided the resulting p value is less than the defined threshold α, the analyst may reject the null hypothesis that the original model is based on random correlations between $\mathbf{X}$ and $\mathbf{Y}$.

3.6.5 CV-ANOVA Testing

Response permutation testing capably reports p values via hypothesis testing of model reliability, but it requires significant computation time to train the 100 – 1,000 models required for an accurate result. The alternative CV-ANOVA testing method effectively requires no additional computation time after internal cross-validation, as it compares the fitted $\mathbf{Y}$ residuals obtained from cross-validation procedures [36]. In effect, CV-ANOVA tests whether the mean square error of fitted residuals from PLS and OPLS models is significantly smaller than the total mean square variation in $\hat{\mathbf{Y}}$. By comparing the ratio of these mean square values to an F distribution, CV-ANOVA reports its own p value which may again be compared to a predefined threshold.

3.7 Conclusions

Multivariate bilinear factorizations such as PCA, PLS and OPLS provide an essential platform for rapid information extraction of rich spectral datasets. Through proper application of processing and treatment, optimal choice of modeling algorithms, and judicious administration of validation metrics, multivariate analysis can lend a powerful hand in biochemical examination of complex, multiparametric metabolic systems. Specific applications of multivariate analysis in metabolomics are discussed in further detail in the following chapter.

3.8 References

[1] K. M. Aberg, E. Alm, and R. J. Torgrip. The correspondence problem for metabonomics datasets. *Analytical and Bioanalytical Chemistry*, 394(1):151–162, 2009.

[2] M. Andersson. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23:518–529, 2009.

[3] M. Baker. Metabolomics: From small molecules to big ideas. *Nature Methods*, 8(2):117–121, 2011.

[4] O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, and J. K. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11):2692–2703, 2007.

[5] J.-C. Boulet and J.-M. Roger. Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61–69, 2012.

[6] L. P. Bras, S. A. Bernardino, J. A. Lopes, and J. C. Menezes. Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour. *Chemometrics and Intelligent Laboratory Systems*, 75(1):91–99, 2005.

[7] G. L. Bretthorst. Bayesian Analysis I. Parameter Estimation Using Quadrature NMR Models. *Journal of Magnetic Resonance*, 88:533–551, 1990.

[8] G. L. Bretthorst. Bayesian Analysis II. Signal Detection and Model Selection. *Journal of Magnetic Resonance*, 88:552–570, 1990.

[9] G. L. Bretthorst. Bayesian Analysis III. Applications to NMR Signal Detection, Model Selection and Parameter Estimation. *Journal of Magnetic Resonance*, 88:571–595, 1990.

[10] G. L. Bretthorst. Nonuniform Sampling: Bandwidth and Aliasing. *Concepts in Magnetic Resonance*, 32A(6):417–435, 2008.

[11] D. E. Brown, T. W. Campbell, and R. N. Moore. Automated Phase Correction of FT NMR Spectra by Baseline Optimization. *Journal of Magnetic Resonance*, 85(1):15–23, 1989.

[12] J. G. Bundy, M. P. Davey, and M. R. Viant. Environmental metabolomics: a critical review and future perspectives. *Metabolomics*, 5(1):3–21, 2009.

[13] J. I. Castrillo, A. Hayes, S. Mohammed, S. J. Gaskell, and S. G. Oliver. An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, 62(6):929–937, 2003.

[14] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis*. Duxbury Press, 1983.

[15] L. Chen, Z. Q. Weng, L. Y. Goh, and M. Garland. An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1-2):164–168, 2002.

[16] D. P. Cherney, D. R. Ekman, D. J. Dix, and T. W. Collette. Raman spectroscopy-based metabolomics for differentiating exposures to triazole fungicides using rat urine. *Analytical Chemistry*, 79(19):7324–7332, 2007.

[17] R. A. Chylla, K. Hu, J. J. Ellinger, and J. L. Markley. Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: Application to quantitative metabolomics. *Analytical Chemistry*, 83(12):4871–4880, 2011.

[18] R. A. Chylla and J. L. Markley. Improved frequency resolution in multidimensional constant-time experiments by multidimensional Bayesian analysis. *Journal of Biomolecular NMR*, 3:515–533, 1993.

[19] R. A. Chylla and J. L. Markley. Theory and application of the maximum likelihood principle to NMR parameter estimation of multidimensional NMR data. *Journal of Biomolecular NMR*, 5(3):245–258, 1995.

[20] R. A. Chylla, B. F. Volkman, and J. L. Markley. Practical model fitting approaches to the direct extraction of NMR parameters simultaneously from all dimensions of multidimensional NMR spectra. *Journal of Biomolecular NMR*, 12(2):277–297, 1998.

[21] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[22] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267, 2006.

[23] E. C. Craig and A. G. Marshall. Automated Phase Correction of FT NMR Spectra by Means of Phase Measurement Based on Dispersion Versus Absorption Relation (DISPA). *Journal of Magnetic Resonance*, 76(3):458–475, 1988.

[24] Q. Cui, I. A. Lewis, A. D. Hegeman, M. E. Anderson, J. Li, C. F. Schulte, W. M. Westler, H. R. Eghbalnia, M. R. Sussman, and J. L. Markley. Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology*, 26(2):162–164, 2008.

[25] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson. Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, 85(1):144–154, 2007.

[26] A. de Juan, Y. Vander Heyden, R. Tauler, and D. L. Massart. Assessment of new constraints applied to the alternating least squares method. *Analytica Chimica Acta*, 346(3):307–318, 1997.

[27] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[28] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[29] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.

[30] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

[31] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, 1998.

[32] W. B. Dunn and D. I. Ellis. Metabolomics: Current analytical platforms and methodologies. *Trends in Analytical Chemistry*, 24(4):285–294, 2005.

[33] H. T. Eastment and W. J. Krzanowski. Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*, 24(1):73–77, 1982.

[34] A. Ebel, W. Dreher, and D. Leibfritz. Effects of zero-filling and apodization on spectral integrals in discrete Fourier-transform spectroscopy of noisy data. *Journal of Magnetic Resonance*, 182:330–338, 2006.

[35] D. I. Ellis and R. Goodacre. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst*, 131(8):875–885, 2006.

[36] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[37] P. Eshghi. Dimensionality choice in principal components analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems*, 130:6–13, 2014.

[38] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, 96(1):22–26, 2009.

[39] D. L. S. Ferreira, S. Kittiwachana, L. A. Fido, D. R. Thompson, R. E. A. Escott, and R. G. Brereton. Window consensus PCA for multiblock statistical process control: adaptation to small and time-dependent normal operating condition regions, illustrated by online high performance liquid chromatography of a three-stage continuous process. *Journal of Chemometrics*, 24(9):596–609, 2010.

[40] J. Forshed, I. Schuppe-Koistinen, and S. P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487(2):189–199, 2003.

[41] T. Gebregiworgis and R. Powers. Application of NMR Metabolomics to Search for Human Disease Biomarkers. *Combinatorial Chemistry and High Throughput Screening*, 15(8):595–610, 2012.

[42] P. Geladi and B. R. Kowalski. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[43] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 4 edition, 2012.

[44] J. Gottfries, E. Johansson, and J. Trygg. On the impact of uncorrelated variation in regression mathematics. *Journal of Chemometrics*, 22:565–570, 2008.

[45] R. Hall, M. Beale, O. Fiehn, N. Hardy, L. Sumner, and R. Bino. Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, 14(7):1437–1440, 2002.

[46] M. Hedenstrom, S. Wiklund, B. Sundberg, and U. Edlund. Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometrics and Intelligent Laboratory Systems*, 92(2):110–117, 2008.

[47] A. Heuer. A New Algorithm for Automatic Phase Correction by Symmetrizing Lines. *Journal of Magnetic Resonance*, 91(2):241–253, 1991.

[48] J. C. Hoch and A. S. Stern. *NMR Data Processing*. Wiley, 1996.

[49] H. C. J. Hoefsloot, M. P. H. Verouden, J. A. Westerhuis, and A. K. Smilde. Maximum likelihood scaling (MALS). *Journal of Chemometrics*, 20(3-4):120–127, 2006.

[50] H. Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2:360–378, 1931.

[51] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[52] D. B. Kell. Metabolomics and systems biology: Making sense of the soup. *Current Opinion in Microbiology*, 7(3):296–307, 2004.

[53] T. Kind, G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn. FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Analytcal Chemistry*, 81(24):10038–10048, 2009.

[54] K. Kjeldahl and R. Bro. Some common misunderstandings in chemometrics. *Journal of Chemometrics*, 24(7-8):558–564, 2010.

[55] P. Koh, E. Chan, M. Mal, K. Eu, A. Blackshall, and H. Keun. Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). *Diseases of the Colon and Rectum*, 52(4):769–769, 2009.

[56] W. J. Krzanowski. Cross-Validation in Principal Component Analysis. *Biometrics*, 43(3):575–584, 1987.

[57] M. H. Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, 2008.

[58] X. Lin, Q. Wang, P. Yun, L. Tang, Y. Tan, H. Li, K. Yan, and G. Xu. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinatorial use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4):549–558, 2011.

[59] J. C. Lindon, J. K. Nicholson, E. Holmes, and J. R. Everett. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance*, 12(5):289–320, 2000.

[60] T. Lofstedt. *OnPLS: Orthogonal Projections to Latent Structures in Multiblock and Path Model Data Analysis*. PhD thesis, Umea University, 2012.

[61] T. Lofstedt and J. Trygg. OnPLS – a novel multiblock method for the modeling of predictive and orthogonal variation. *Journal of Chemometrics*, 25:441–455, 2011.

[62] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Uncorrelated Multilinear Principal Component Analysis for Unsupervised Multilinear Subspace Learning. *IEEE Transactions on Neural Networks*, 20(11):1820–1836, 2009.

[63] H. P. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, 2011.

[64] R. Manne. Analysis of two partial least squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 1:187–197, 1987.

[65] D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015.

[66] E. M. S. McNiven, J. B. German, and C. M. Slupsky. Analytical metabolomics: nutritional opportunities for personalized health. *Journal of Nutritional Biochemistry*, 22(11):995–1002, 2011.

[67] M. Mobli and J. C. Hoch. Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 83C:21–41, 2014.

[68] N. P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17–35, 1998.

[69] R. Powers. NMR metabolomics and drug discovery. *Magnetic Resonance in Chemistry*, 47:S2–S11, 2009.

[70] R. Ramautar, A. Demirci, and G. J. de Jong. Capillary electrophoresis in metabolomics. *Trends in Analytical Chemistry*, 25(5):455–466, 2006.

[71] S. S. Rubakhin, E. V. Romanova, P. Nemes, and J. V. Sweedler. Profiling metabolites and peptides in single cells. *Nature Methods*, 8(4):S20–S29, 2011.

[72] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[73] A. D. Schuyler, M. W. Maciejewski, A. S. Stern, and J. C. Hoch. Formalism for hypercomplex multidimensional NMR employing partial-component subsampling. *Journal of Magnetic Resonance*, 227:20–24, 2013.

[74] J. Shao. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[75] M. M. Siegel. The Use of the Modified Simplex-Method for Automatic Phase Correction in Fourier-Transform Nuclear Magnetic-Resonance Spectroscopy. *Analytica Chimica Acta*, 5(1):103–108, 1981.

[76] A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. J. C. van der Werff-van der Vat, and R. H. Jellema. Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77(20):6729–6736, 2005.

[77] A. K. Smilde, J. A. Westerhuis, and S. de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337, 2003.

[78] S. A. A. Sousa, A. Magalhaes, and M. M. C. Ferreira. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122:93–102, 2013.

[79] A. D. Southam, T. G. Payne, H. J. Cooper, T. N. Arvanitis, and M. R. Viant. Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry*, 79(12):4595–602, 2007.

[80] D. J. States, R. A. Haberkorn, and D. J. Ruben. A Two-Dimensional Nuclear Overhauser Experiment with Pure Absorption Phase in Four Quadrants. *Journal of Magnetic Resonance*, 48:286–292, 1982.

[81] G. Stoch and Z. Olejniczak. Missing first points and phase artifacts are mutually entangled in FT NMR data – noniterative solution. *Journal of Magnetic Resonance*, 173(2):140–152, 2005.

[82] J. Tang. Microbial metabolomics. *Current Genomics*, 12(6):391–403, 2011.

[83] H. S. Tapp and E. K. Kemsley. Notes on the practical utility of OPLS. *Trends in Analytical Chemistry*, 28(11):1322–1327, 2009.

[84] B. H. ter Kuile and H. V. Westerhoff. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters*, 500(3):169–171, 2001.

[85] G. Tomasi, F. Savorani, and S. B. Engelsen. Icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A*, 1218(43):7832–7840, 2011.

[86] R. J. O. Torgrip, K. M. Aberg, E. Alm, I. Schuppe-Koistinen, and J. Lindberg. A note on normalization of biofluid 1D ^{1}H NMR data. *Metabolomics*, 4(2):114–121, 2008.

[87] R. N. Trethewey. Gene discovery via metabolic profiling. *Current Opinion in Biotechnology*, 12(2):135–138, 2001.

[88] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[89] J. Trygg and S. Wold. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, 17(1):53–64, 2003.

[90] S. Tyagi, Raghvendra, U. Singh, T. Kalra, and K. Munjal. Applications of Metabolomics – a systematic study of the unique chemical fingerprints: an overview. *International Journal of Pharmaceutical Sciences Review and Research*, 3(1):83–86, 2010.

[91] R. a. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(142):1–15, 2006.

[92] K. A. Veselkov, J. C. Lindon, T. M. D. Ebbels, D. Crockford, V. V. Volynkin, E. Holmes, D. B. Davies, and J. K. Nicholson. Recursive Segment-Wise Peak Alignment of Biological ^{1}H NMR Spectra for Improved Metabolic Biomarker Recovery. *Analytical Chemistry*, 81(1):56–66, 2009.

[93] N. Vinayavekhin, E. A. Homan, and A. Saghatelian. Exploring Disease through Metabolomics. *ACS Chemical Biology*, 5(1):91–103, 2010.

[94] V. N. Viswanadhan, H. Rajesh, and V. N. Balaji. Atom Type Preferences, Structural Diversity, and Property Profiles of Known Drugs, Leads, and Nondrugs: A Comparative Assessment. *ACS Combinatorial Science*, 13(3):327–336, 2011.

[95] W. Weckwerth. Metabolomics in systems biology. *Annual Review of Plant Biology*, 54:669–689, 2003.

[96] J. A. Westerhuis and P. M. J. Coenegracht. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11(5):379–392, 1997.

[97] J. A. Westerhuis, S. de Jong, and A. K. Smilde. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 56(1):13–25, 2001.

[98] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[99] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[100] J. A. Westerhuis, E. J. J. van Velzen, H. C. J. Hoefsloot, and A. K. Smilde. Discriminant Q^2 (DQ^2) for improved discrimination in PLSDA models. *Metabolomics*, 4(4):293–296, 2008.

[101] K. M. Wilcoxen, T. Uehara, K. T. Myint, Y. Sato, and Y. Oda. Practical metabolomics in drug discovery. *Expert Opinions in Drug Discovery*, 5(3):249–263, 2010.

[102] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35:521–526, 2007.

[103] S. Wold, E. Johansson, and M. Cocchi. *PLS: Partial Least Squares Projections to Latent Structures*. KLUWER ESCOM Science Publisher, 1993.

[104] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[105] K. Wongravee, N. Heinrich, M. Holmbee, M. Schaefer, R. R. Reed, J. Trevejo, and R. G. Brereton. Variable Selection Using Iterative Reformulation of Training Set Models for Discrimination of Samples: Application to Gas Chromatography/Mass Spectrometry of Mouse Urinary Metabolites. *Analytical Chemistry*, 81(13):5204–5217, 2009.

[106] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013.

[107] B. Worley and R. Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

[108] B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014.

[109] W. Wu, M. Daszykowski, B. Walczak, B. C. Sweatman, S. C. Connor, J. N. Haseldeo, D. J. Crowther, R. W. Gill, and M. W. Lutz. Peak alignment of urine NMR spectra using fuzzy warping. *Journal of Chemical Information and Modeling*, 46(2):863–875, 2006.

[110] Q. S. Xu and Y. Z. Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

[111] B. Zhang, S. Halouska, R. Gaupp, S. Lei, E. Snell, R. J. Fenton, R. G. Barletta, G. A. Somverville, and R. Powers. Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes. *Journal of Integrated OMICS*, 2(3):120–137, 2013.

[112] M. Zhou, J. F. McDonald, and F. M. Fernandez. Optimization of a direct analysis in real time/time-of-flight mass spectrometry method for rapid serum metabolomic fingerprinting. *Journal of the American Society for Mass Spectrometry*, 21(1):68–75, 2010.

Chapter 4
Applications of Multivariate Analysis in Metabolomics

4.1 Introduction

This chapter details several varied applications of multivariate analysis within the field of metabolomics, from the simplest examples of constructing multivariate calibration models of ^{1}H NMR spectral data for determination of caffeine concentration in coffee, to more complex examples of multiblock statistical modeling of joint ^{1}H NMR and electrospray MS data. A final note on the relationship between PCA scores-space class separations and OPLS-DA model reliability is also presented to conclude the chapter.

4.2 ^{1}H NMR Fingerprinting of Brewed Coffees

To provide an initial illustration of the capabilities of the MVAPACK software suite [32], four roasts of brewed coffee were purchased from a local coffee shop. In this study, ^{1}H NMR and UV/Vis absorbance spectra were collected in order to construct a multivariate calibration of ^{1}H NMR spectral information against caffeine concentration.

4.2.1 Materials and Methods

Coffee Sample Preparation

Four freshly brewed roasts of coffee (Light, Dark, Medium Regular and Medium Decaffeinated) were purchased from a local coffee shop. From each roast, sixteen 1.2 mL samples were drawn while the coffee was still hot and stored at $-80°$C for 24 hours. The samples were then lyophilized at $-50°$C and 0.1 mBar for 24 hours and subsequently redissolved in 1.0 mL of 99.8% D_2O (Isotec, St. Louis, MO) without pH adjustment. Following re-dissolution, the samples were centrifuged at 12,000 RPM and 25°C for 5 minutes and 800 μL of the supernatant was collected into NMR tubes. The samples were stored in their NMR tubes at 4°C for 36 hours prior to data collection.

Caffeine Extraction

Measurement of the caffeine concentration in each coffee roast was performed based on previously outlined procedures [3]. Triplicate standards of caffeine were made by dissolving 2.9 mg of caffeine (Sigma-Aldrich, St. Louis, MO) into 100.0 mL of 99.5% CH_2Cl_2 (Sigma-Aldrich, St. Louis, MO) for a final concentration of 149 μM. From each purchased coffee roast, 25 mL of brewed coffee were combined with 25 mL of CH_2Cl_2 in a separatory funnel in a two-step liquid-liquid extraction. Extracted caffeine in CH_2Cl_2 was diluted 20-fold into 1.0 mL and subjected to UV/Vis absorption spectroscopy for caffeine quantitation.

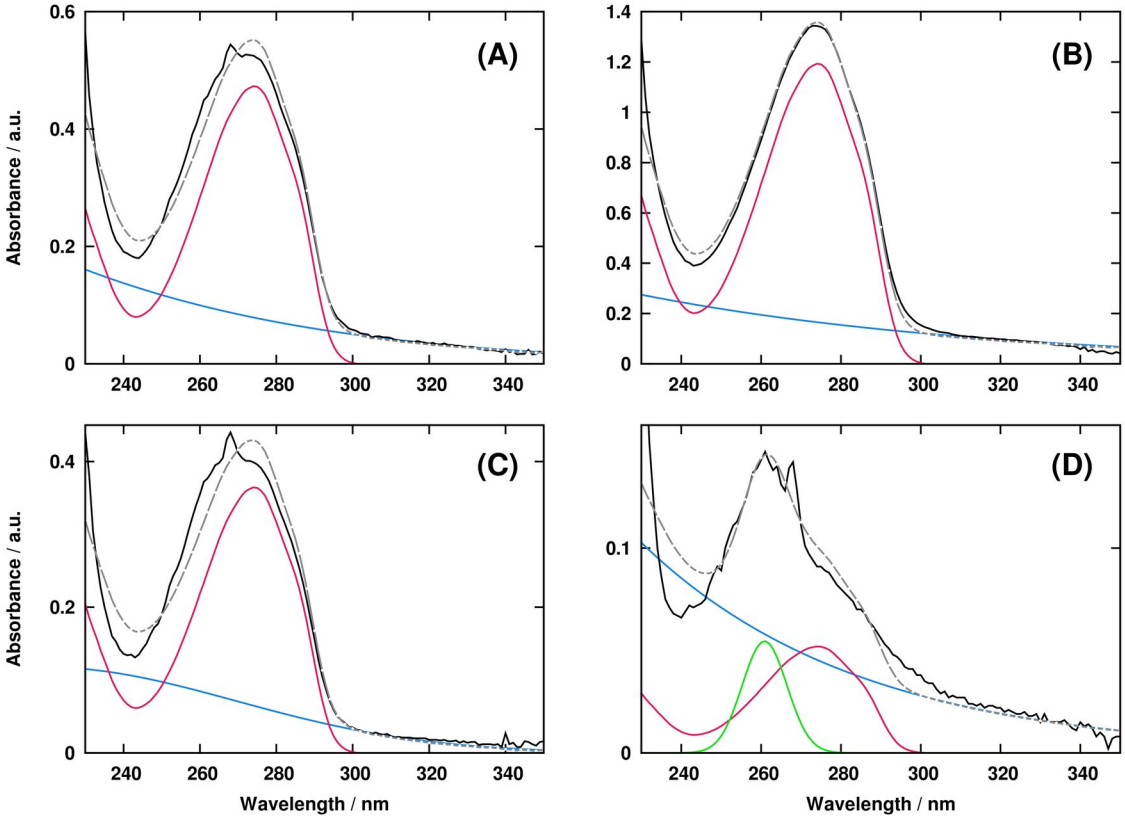

Figure 4.1: UV/Vis Caffeine Quantitation Band-fitting Results.
UV/Vis absorbance band-fitting results for caffeine concentration estimation of dark roast (**A**), light roast (**B**), regular medium roast (**C**), and decaffeinated medium roast (**D**). Black lines represent observed spectra, dashed grey lines represent fitted spectra, red lines represent fitted caffeine, and blue and green lines represent additional Gaussian bands required for fitting.

UV/Vis Spectroscopy

Absorption spectra of caffeine standards and extracts were collected on a Shimadzu UV-2501PC with a 1.0 nm slit width and 1.0 cm quartz cuvettes. Spectra were collected between the wavelengths of

500 nm and 230 nm.

NMR Spectroscopy

All NMR experiments were collected on a Bruker Avance DRX 500 MHz spectrometer equipped with a 5 mm inverse triple-resonance (^{1}H, ^{13}C, ^{15}N) cryoprobe with a z-axis gradient. A Bruker BACS-120 sample changer and ICON-NMR software were used to automate NMR data collection. A standard 1D ^{1}H NMR spectrum using a SOGGY pulse sequence [14, 19] and a T_2-filtered 1D ^{1}H NMR spectrum using a z-filtered Carr-Purcell-Meiboom-Gill (CPMG) sequence [20] with an identical SOGGY water suppression element were acquired for each sample. All experiments were performed at 20°C with 128 scans, 32 dummy scans, a carrier frequency offset of 2,351 Hz, a 6,009 Hz spectral width, and a 1.0 s inter-scan delay. For T_2 filtered spectra, 20 repetitions of a CPMG-z element having a delay (τ) of 5.0 ms were performed per scan, for a total filter time ($2n\tau$) of 200.0 ms. Free induction decays were collected with 32,768 total data points resulting in a total acquisition time of 10 minutes per experiment.

Caffeine Quantitation

A reference spectrum of caffeine in CH_2Cl_2 was generated from the three standard UV/Vis absorption spectra by taking the mean of the spectra after multiplicative scatter correction (MSC, [12]). To quantify caffeine in the extracts, the absorption spectrum of each extract was fit by nonlinear least squares [16] to the sum of the scaled caffeine reference spectrum and no more than two extra "background" Gaussian bands (Figure 4.1). The ratio of the fit caffeine reference spectrum in each extract to that of the known samples was used as an estimate of caffeine concentration in the extracts. Concentrations of the medium regular, medium decaffeinated, dark and light roasts were 1.526 mM, 0.217 mM, 1.979 mM and 4.993 mM, respectively.

Multivariate Analysis

All NMR spectra were loaded, processed, treated and modeled inside the GNU Octave 3.6 programming environment [9] using functions available in the MVAPACK software suite for chemometrics [32]. Free induction decays were loaded in from Bruker DMX binary format and corrected for group delay errors by a circular shift of their time-domain data points. All decays were Fourier transformed,

automatically phase-corrected and referenced to match the chemical shifts of caffeine with known database values. Spectral regions upfield of 0.44 ppm and downfield of 9.16 ppm were removed from the dataset, as they contained no informative signals. As solvent resonances were adequately suppressed by the excitation sculpting pulse sequence, no spectral regions were removed around the water resonance. Figure 4.2 illustrates the final result of spectral processing of the coffees dataset using MVAPACK.

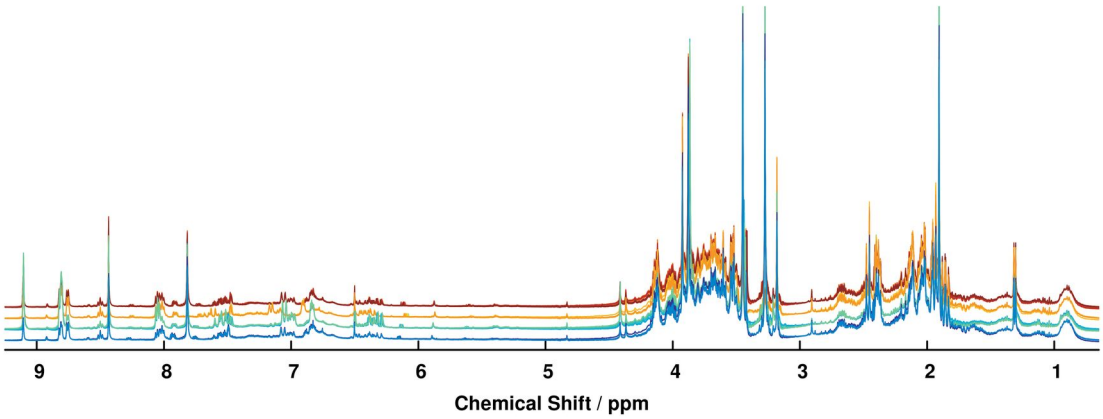

Figure 4.2: Processed ^{1}H NMR Spectra of Coffee Roasts.
Representative processed 1D ^{1}H NMR spectra for all spectra of each coffee roast, acquired using the water-suppressed CPMG-z pulse sequence and processed in MVAPACK. To reach this point, free induction decays were simply Fourier transformed and automatically phased. No manual phase corrections were applied after autophasing.

For principal component analysis (PCA), the dataset was normalized by the method of probabilistic quotients (PQ, [8]) and subjected to adaptive intelligent binning [7]. Low-variation bins were automatically removed from the dataset [37], resulting in a final data matrix having 64 observations and 284 variables. The data matrix was scaled to unit variance [25] prior to NIPALS PCA modeling [15], which produced six significant components having cumulative R^2_X and Q^2 statistics of 0.9689 and 0.8965 ± 0.0105, respectively [11].

Linear discriminant analysis (LDA) was performed on the first three dimensions of resulting PCA scores to yield a two-component model that best captured the between-class variation present in the three orthonormal PCA score vectors. LDA modeling yielded a model having a cumulative R^2_X statistic of 0.9950 and cumulative R^2_Y and Q^2 statistics of 1.0. Scores from the PCA model of the coffees ^{1}H NMR spectral data, and their corresponding LDA projection, are shown in Figure 4.3.

For orthogonal projections to latent structures regression (OPLS-R, [23]), the full-resolution dataset

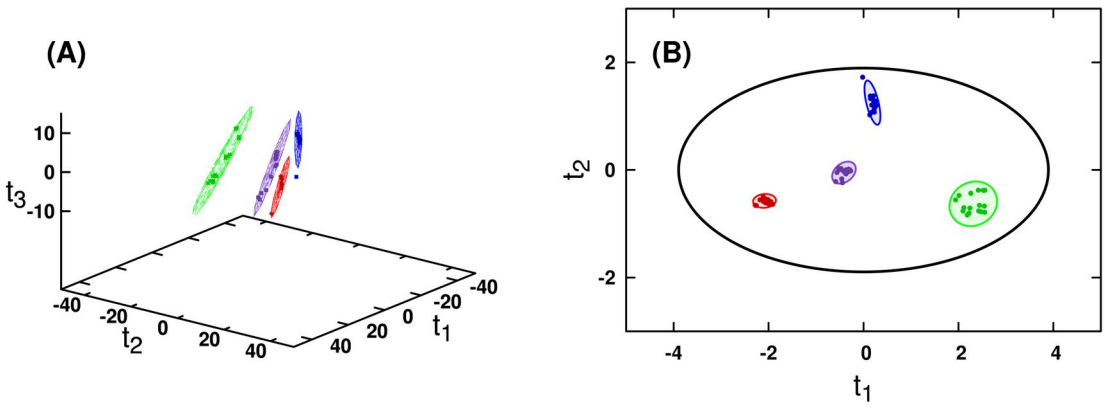

Figure 4.3: Principal Component Scores of the Coffees Spectra.
PCA (**A**) and LDA (**B**) scores of the four coffee roasts. Red, green, blue and violet points represent dark, light, decaffeinated medium, and regular medium roasts, respectively. Ellipsoids and ellipses enclose the 95% confidence intervals estimated by the sample means and covariances of scores from each class. Axis labels in panels (**A**) and (**B**) indicate scores in PCA and LDA bases, respectively, and not the same set of scores.

was aligned using a per-class application of interval correlation-optimized shifting (*i*COshift, [21]) and PQ normalization, resulting in a final data matrix having 64 observations and 11,888 variables. The Pareto-scaled data matrix was regressed by OPLS against a response vector containing caffeine concentrations estimated by UV/Vis analysis of the four coffee roasts, yielding a model with one predictive component and one orthogonal component ($R^2_{X,p} = 0.5294$, $R^2_{X,o} = 0.1288$, $R^2_Y = 0.9822$, $Q^2 = 0.9502 \pm 0.0008$). CV-ANOVA significance testing returned a p value equal to zero ($F = 2258.8$) to within double-precision floating point error, indicating a reliable model. The OPLS-R and LDA models were further validated using response permutation tests having 1,000 iterations each. The permutation tests of both models resulted in p values less than 0.001 for both R^2_Y and Q^2, a further indication of high model reliability.

Validation against SIMCA-P+

Correctness of the PCA and OPLS-R models generated by MVAPACK was verified by exporting the final processed and treated data matrices from GNU Octave and modeling them in SIMCA-P+ 13.0 (Umetrics AB, Umeå, Sweden). The scores extracted from SIMCA and MVAPACK were found to have coefficients of determination (R^2) of 0.999976 and 0.999989 for the PCA and OPLS models, respectively. The "imperfect" non-unity values of R^2 reflect the fact that SIMCA-P+ 13.0 only permits the export of scores with no more than four decimal places.

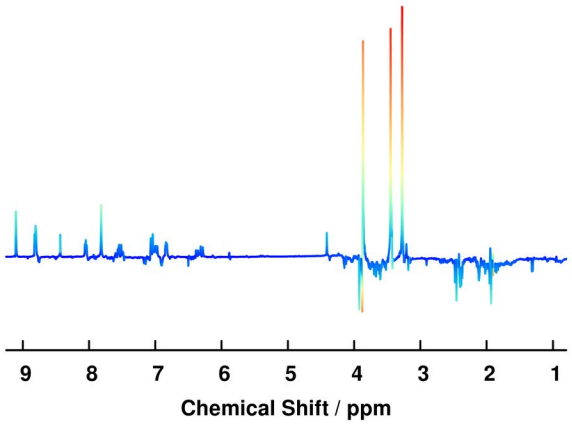

Figure 4.4: Backscaled Coffees OPLS-R Model Loadings. Backscaled OPLS-R predictive loadings of the four coffee roasts regressed according to estimated caffeine concentration. The pseudospectral nature of backscaled loadings facilitates analysis of model results by any spectroscopist. The four most intense positive positive peaks in the loadings pseudospectrum correspond directly to caffeine NMR resonances archived in the BMRB, indicating a fairly successful regression against caffeine concentration.

4.2.2 Results and Discussion

Use of MVAPACK during analysis of the coffees dataset arguably facilitated rapid identification of ideal processing, treatment and modeling parameters during data handling. Use of automatic phase correction, adaptive intelligent binning, and PQ normalization yielded a dataset in which three principal components were sufficient to fully separate all classes in scores space, and subsequent LDA modeling resulted in complete class separation in only two components (Figure 4.3).

As opposed to the PCA modeling procedure, which utilized binned spectra, OPLS-R model training was performed using full-resolution 1D ^{1}H NMR spectra in order to reap the interpretive advantages of full-resolution backscaled loadings (Figure 4.4). The availability of *i*COshift alignment [21] in MVAPACK effectively makes the modeling of full-resolution NMR spectral data possible by correcting positional noise [1] in the spectra that corrupts the bilinear nature of the data. By regressing the NMR data against estimates of caffeine concentration obtained by UV/Vis spectroscopy (Figure 4.1), a loading pseudo-spectrum of caffeine was obtained that matched almost perfectly with spectral data deposited in the Biological Magnetic Resonance Bank [24]. It is conceivable that spectral features co-extracted with caffeine in the loadings correspond to coffee bean metabolites lost alongside caffeine during roasting or decaffeination.

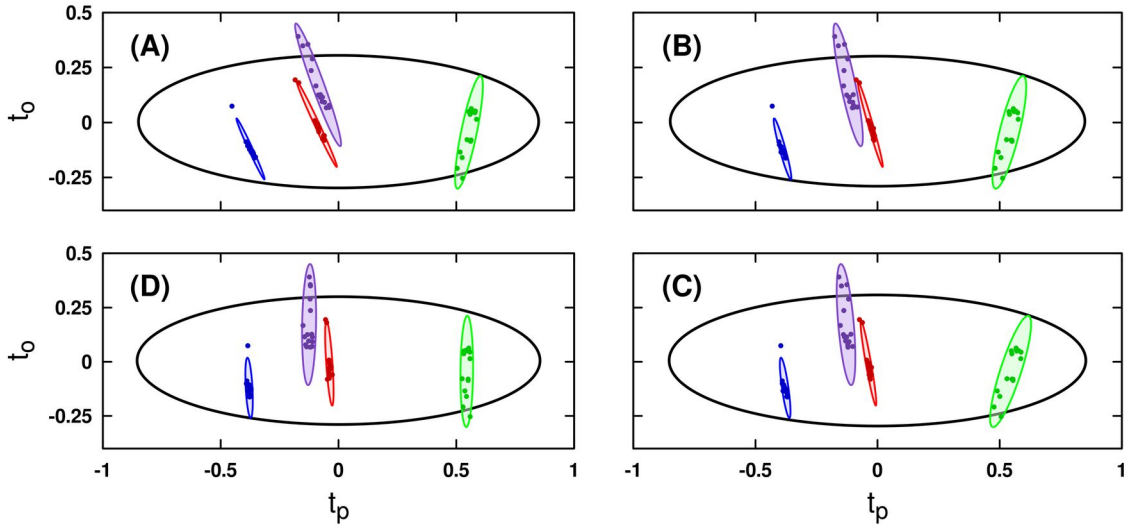

Figure 4.5: Coffees OPLS-R Scores as Evidence of Overfit.
OPLS-R scores of the four coffee roasts, where each roast was regressed against its caffeine concentration estimated by UV/Vis absorbance spectroscopy. Scores in panels (**A**) through (**D**) were computed from models having 1 through 4 orthogonal components, respectively.

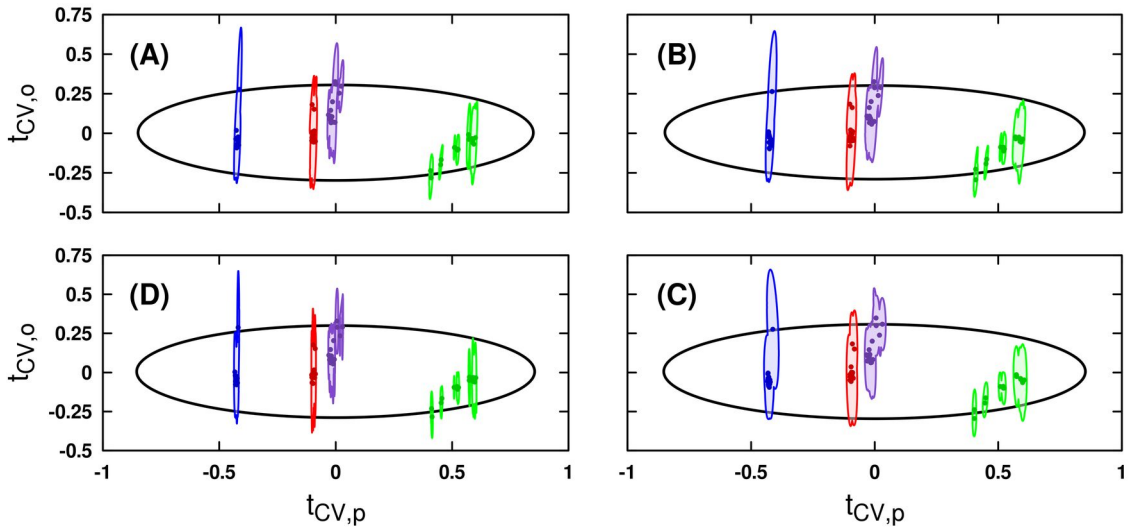

Figure 4.6: Coffees OPLS-R Cross-validated Scores.
OPLS-R scores of the four coffee roasts, where each roast was regressed against its caffeine concentration, as in Figure 4.5. Mean score values and confidence ellipses for each observation were computed from 100 iterations of seven-fold Monte Carlo internal cross-validation.

Notably, the UV/Vis-estimated caffeine concentration of the dark roast coffee was slightly higher than that of the medium roast, which is contrary to expectation given that the coffees were brewed using equal volumes of grounds. However, OPLS-R of the NMR data using the estimated caffeine concentrations correctly ranked the roasts according to expectation (Figure 4.5A). When more or-

thogonal components were allowed into the OPLS-R model, the dark roast again shifted to a higher caffeine concentration, beautifully indicating the presence of overfitting (Figure 4.5B–D). Monte Carlo cross-validated scores further supported the fact that a $1 + 1$ OPLS model was the most appropriate (Figure 4.6). Therefore, an OPLS-R model having only a single orthogonal component was chosen, given the fact that it more faithfully modeled the underlying NMR data at the expense of contradicting the more uncertain UV/Vis measurements.

Finally, no discernible difference was observed between the 1D ^{1}H NMR spectra acquired with and without T_2-filtering. Spectra collected on in-house brewed coffee exhibited high levels of protein background signal, which were readily suppressed using the CPMG-z pulse sequence element. On the other hand, the spectra of the four purchased roasts showed no such background signal, possibly due to more correct brewing technique.

4.3 Fingerprinting of Joint ^{1}H NMR and DI-ESI-MS Data

Multiblock bilinear factorizations such as CPCA-W, MB-PLS and MB-OPLS provide a powerful framework for analyzing a set of multivariate observations from multiple analytical measurements containing potentially correlated variables [26, 28, 22]. Such algorithms provide analogous information to PCA, PLS and OPLS in situations where extra knowledge is available to subdivide the measured variables into multiple "blocks". As a result, the correlation structures of each block *and* the between-block correlations may be simultaneously utilized. Due to the existence of common trends among all blocks, this use of between-block correlations during modeling will ideally bring the model loadings (latent variables) into better agreement with the true underlying biochemistry (hidden variables). In short, multiblock algorithms provide an ideal means of integrating 1D ^{1}H NMR and direct injection electrospray mass spectrometry (DI-ESI-MS) datasets for metabolic fingerprinting [35].

Consensus PCA (CPCA-W), Multiblock PLS (MB-PLS), and Multiblock OPLS (MB-OPLS) were used to analyze 1D ^{1}H NMR and DI-ESI-MS data collected on metabolite extracts from human dopaminergic neuroblastoma cells (SK-N-SH) after different neurotoxin treatments [17]. Each dataset was also individually subjected to single-block modeling by PCA and PLS in order to highlight the information gained by jointly modeling the data within multiblock frameworks.

4.3.1 Materials and Methods

NMR Acquisition and Processing

NMR data were collected and processed according to previously described procedures [36]. A Bruker Avance DRX 500 MHz spectrometer equipped with a 5 mm inverse triple-resonance cryoprobe (^{1}H, ^{13}C, ^{15}N) with a z-axis gradient, a BACS-120 sample changer, and an automatic tuning and matching accessory were utilized for automated NMR data collection. Free induction decays were collected into 32,768 complex data points over a spectral window of 2,342±2,741 ppm, using the SOGGY water suppression pulse sequence (*zgesgp*, [14, 19]).

Following acquisition, the 1D ^{1}H NMR free induction decays were processed in the MVAPACK toolbox [32]. A 1.0 Hz exponential apodization function and a single round of zero-filling were applied prior to Fourier transformation. Spectra were then automatically phased and normalized using phase-scatter correction (PSC, [33], Chapter 6). Finally, chemical shift regions containing spectral baseline noise or solvent signals were manually removed. Binning of the processed NMR spectra was performed using the Adaptive Intelligent (AI) binning algorithm that avoids splitting signals into multiple bins [7].

MS Acquisition and Processing

Mass spectra of the SK-N-SH metabolite extracts were acquired in positive ion mode over a mass range of m/z 50–1,200. Spectra were acquired for 30 s each using the following source conditions: 2.5 kV electrospray capillary voltage, 60 V sampling cone voltage, 4.0 V extraction voltage, 80°C source temperature, 250°C desolvation temperature, 500 L/h desolvation gas flow rate, and 15 μL/min injection flow rate.

The initial stages of mass spectral data processing were performed using MassLynx V4.1 (Waters Corp., Milford, MA). A background subtraction was performed on all spectra: reference spectra of either paraquat, 1-methyl-4-phenylpyridinium (MPP^{+}), rotenone, or 6-hydroxydopamine (6-OHDA) in $H_2O/CH_3OH/HCO_2H$ (49.75:49.75:0.5) at 10 ppm were used as backgrounds. Background subtraction of each spectrum was performed in a class-dependent manner (e.g. the MPP^{+} reference mass spectrum was used as background for MPP^{+}-treated cell samples). As a result, mass spectral signals from the drugs themselves were guaranteed to not influence subsequent analyses. The

background-subtracted mass spectra were then loaded into MVAPACK for binning and normalization. All mass spectra were linearly re-interpolated onto a common axis that spanned from m/z 50–1,200 in 0.003 m/z steps, resulting in 383,334 variables prior to processing. Based on the low probability of observing a metabolite in the mass range m/z 1,100–1,200, the region was removed prior to binning. Mass spectra were uniformly binned using a bin width of 0.5 m/z, resulting in a data matrix of 2,095 variables. Finally, the MS data matrix observations were normalized using probabilistic quotient (PQ) normalization [8].

Multivariate Statistical Analysis

Using functions available in the latest version of MVAPACK, the NMR and MS data were joined into a single multiblock data structure and modeled using CPCA-W, MB-PLS and MB-OPLS. Both blocks were scaled to unit variance prior to modeling, and equal contribution of each block to the models (fairness) was ensured by further scaling each block by the square root of its variable count [22]. For the purposes of comparison, PCA and PLS models of the independent NMR and MS data matrices were also constructed. All PLS models were trained on a binary discriminant response matrix (i.e. PLS-DA), in which untreated cells were assigned to one class, and all neurotoxin-treated cells were assigned to a second class.

Cross-validation of Multivariate Models

Initially, all PCA and CPCA-W models were internally cross-validated using a leave-one-out (LOOCV) procedure in MVAPACK during model training [11]. A subsequent set of PCA models was trained and cross-validated using a Monte Carlo seven-fold (MCCV) procedure that produced less optimistic Q^2 statistics. All PLS-DA, MB-PLS-DA and MB-OPLS-DA models were internally cross-validated using a Monte Carlo seven-fold procedure [30]. All MCCV rounds involved 50 iterations per tested model component. The results of cross-validation were summarized by per-component Q^2 statistics, and the number of model components was chosen such that the cumulative Q^2 was a strictly increasing function of component count. Response permutation tests of all supervised models were performed with 1,000 permutations each to assess the statistical significance of R^2_Y and Q^2 values [27]. CV-ANOVA significance tests [10] were also performed to supplement the results of the permutation tests.

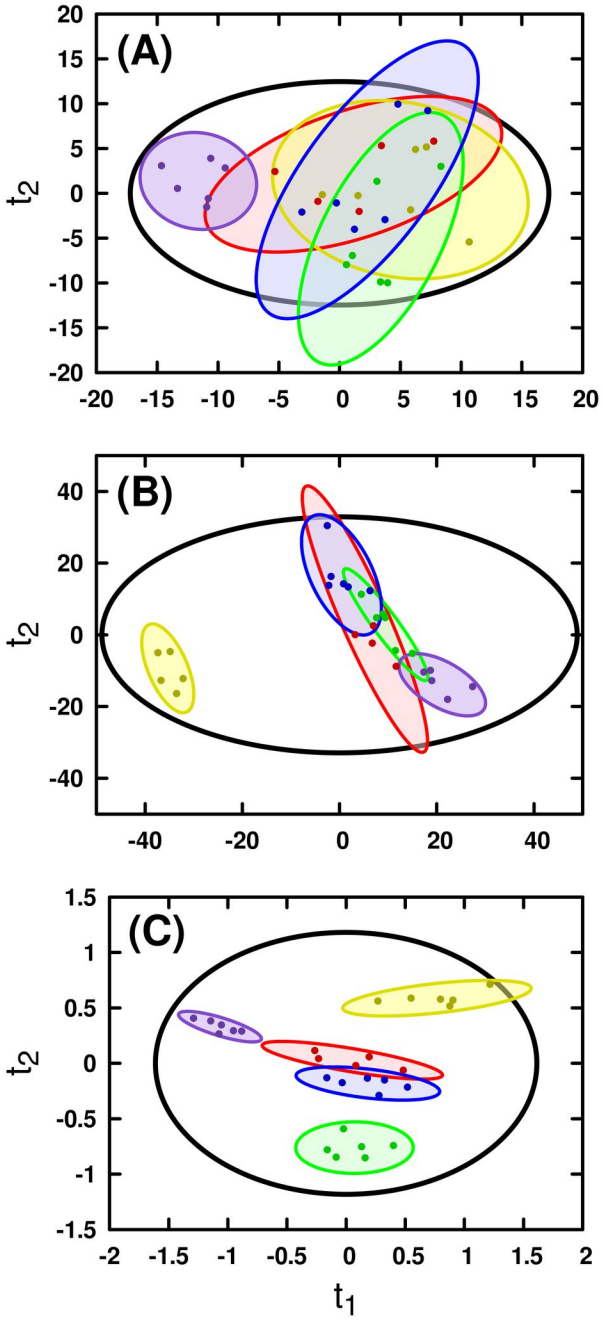

Figure 4.7: Comparison of PCA and MB-PCA Scores.
Scores generated from (**A**) PCA of ^{1}H NMR in vacuo, (**B**) PCA of DI-ESI-MS in vacuo, and (**C**) MB-PCA of ^{1}H NMR and DI-ESI-MS. Separations between classes are increased upon combination of the two data matrices via MB-PCA. Yellow, red, green, violet and blue scores correspond to the control, 6-OHDA, MPP^{+}, paraquat and rotenone classes, respectively.

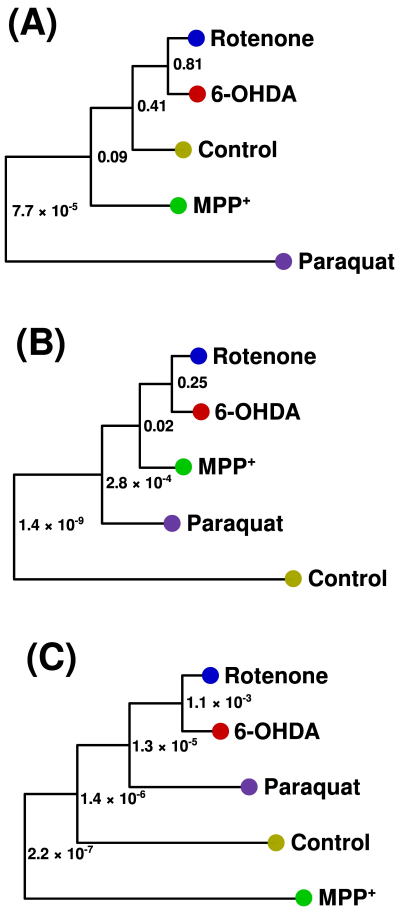

Figure 4.8: Dendrograms of PCA and MB-PCA Scores. Dendrograms computed from scores-space class separations [31] in the in vacuo PCA and MB-PCA models. Panels (**A**–**C**) correspond to scores in panels (**A**–**C**) in Figure 4.7, above.

4.3.2 Results and Discussion

Classical Modeling

PCA of the binned NMR data matrix ($N = 29$, $K = 159$) resulted in 10 principal components having cumulative R^2_X and Q^2 statistics of 0.9485 and 0.4591, respectively, based on LOOCV. Overall, no patterns were readily discernible in the NMR PCA scores (Figure 4.7A) due to high within-class variation in the data. However, scores for paraquat treatment were found to significantly separate from all other classes ($p < 0.002$) along the first principal component. Scores from PCA of the binned MS data matrix ($N = 29$, $K = 2{,}095$) were found to exhibit markedly less within-class variation compared to the NMR data (Figure 4.7B). Using LOOCV, three significant components were identified from the binned MS data, yielding fairly low cumulative R^2_X and Q^2 statistics of 0.3397 and 0.1590. While paraquat treatment still separated from other drug treatments in MS PCA scores space, the greatest separations were observed between treated and untreated (control) cells ($p < \times 10^{-9}$). These differing patterns of separation in NMR and MS PCA scores suggested that multiblock analyses could provide further information, ideally separating both control and paraquat scores from all other classes. Figure 4.8 contains dendrograms of scores-space class separation for each scores plot in Figure 4.7.

Per-component Q^2 statistics computed from LOOCV of the NMR and MS PCA models suggested fairly marginal model reliability at component counts greater than one, so follow-up analyses were performed using MCCV on the same data matrices to obtain less optimistic estimates of reliability. In both cases, MCCV produced single-component PCA models, indicating that the LOOCV had substantially over-estimated the number of principal components in each matrix. A comparison of the resulting Q^2 statistics from LOOCV and MCCV is shown in Figure 4.9.

PLS-DA of the full-resolution NMR ($N = 29$, $K = 16{,}138$) and MS ($N = 29$, $K = 2{,}095$) data matrices both resulted in two-component models. With the exception of the algorithmically forced separation between control and treatment classes, similar clustering patterns were observed when compared to the PCA scores (Figure 4.10A–B). MCCV results from the NMR ($R^2_Y = 0.9519, Q^2 = 0.7303 \pm 0.0517$) and MS ($R^2_Y = 0.9951, Q^2 = 0.9440 \pm 0.01142$) PLS-DA models indicated reasonable levels of response fit and predictive ability. Further validation by CV-ANOVA [10] indicated reliable models with p values of 0.002 and 9.8×10^{-6} for NMR and MS data, respectively. Response permutation tests for both PLS-DA models returned $p < 0.001$, supporting the CV-ANOVA test

results.

Multiblock Modeling

Identification of consensus directions in the NMR and MS data matrices that maximally captured overall variation (MB-PCA) or response correlations (MB-PLS) resulted in more informative models than those calculated against either NMR or MS in vacuo. Using MB-PCA with LOOCV, five significant components were identified ($Q^2 = 0.2322$) that cumulatively explained comparable amounts of variation in the NMR ($R^2_X = 0.8528$) and MS ($R^2_X = 0.5015$) blocks relative to the individual PCA models. As expected, MB-PCA combined the information from both blocks to dramatically increase class separations in super-scores space (Figure 4.7C). More specifically, both control and paraquat classes were separated from other neurotoxin treatments, predominantly along the first principal component. Furthermore, MPP^+ treatment exhibited significant separation from 6-OHDA and rotenone treatments, which was not expected from examination of the individual NMR or MS PCA scores.

Subsequent re-evaluation of the MB-PCA model's reliability using MCCV reduced the number of expected significant components to two (Figure 4.9C). However, secondary and higher components' Q^2 statistics were still found to be statistically indistinguishable from zero, which further suggests that only one principal direction exists in the binned NMR and MS data matrices that captures any substantial variation.

MB-PLS of the data yielded similar improvements in model information content. Two significant components were identified ($R^2_Y = 0.9876, Q^2 = 0.9014 \pm 0.0185$) that clearly separated control and paraquat treatment classes from all other classes in scores space (Figure 4.10). CV-ANOVA testing produced a p value of 3.4×10^{-4} and response permutation testing yielded $p < 0.001$, indicating a reliable MB-PLS-DA model. Backscaled first-component MB-PLS-DA loadings are shown in Figure 4.11. Modeling the multiblock data with MB-OPLS and MCCV produced a single predictive component and a single orthogonal component ($R^2_Y = 0.9031, Q^2 = 0.7084 \pm 0.0241$), making later interpretation markedly simpler (cf. Figures 4.12 and 4.13). Examination of cross-validated MB-OPLS-DA scores (Figure 4.14) provides an excellent example of how PLS mixes predictive and compensatory variation. In MB-OPLS super-scores, paraquat treatment is distinctly separated from other neurotoxin treatment classes along the orthogonal component ($\mathbf{t_o}$). In the MB-PLS model, this distinction between paraquat and other drug treatments becomes mixed with the variation that separates the control class from all drug treatments. However, the two effects have been disentangled

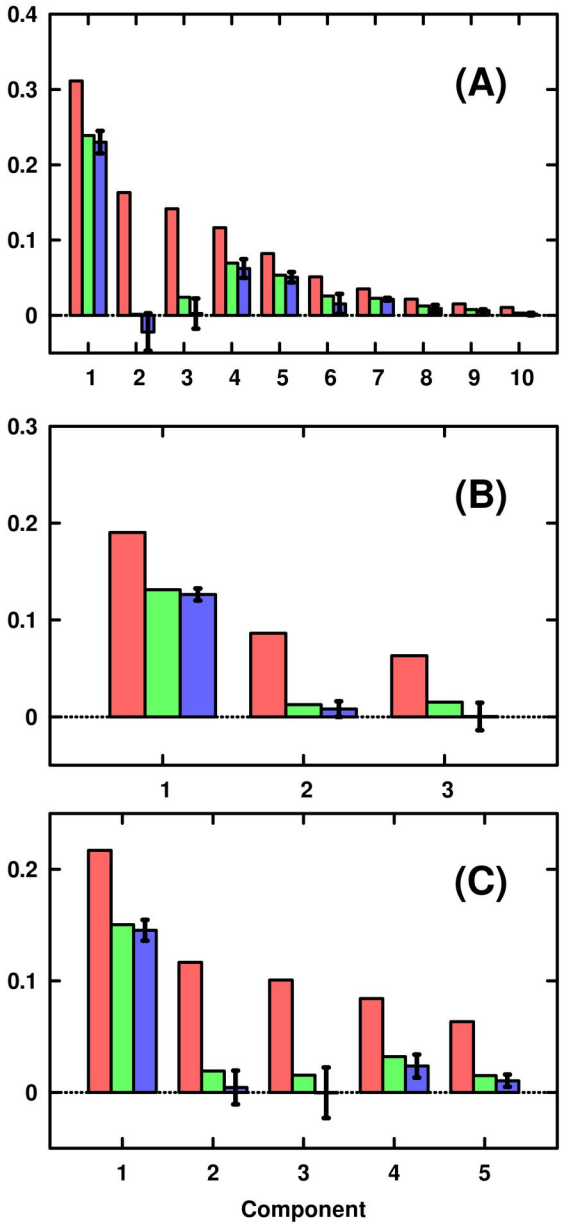

Figure 4.9: Comparison of LOOCV and MCCV Q^2 Statistics for PCA. R^2 (red), Q^2_{LOOCV} (green) and Q^2_{MCCV} (blue) statistics from (**A**) PCA of ^{1}H NMR in vacuo, (**B**) PCA of DI-ESI-MS in vacuo, and (**C**) MB-PCA of ^{1}H NMR and DI-ESI-MS. In all cases, MCCV indicates that both datasets contain a single significant principal component, while LOOCV overestimates the number of significant components.

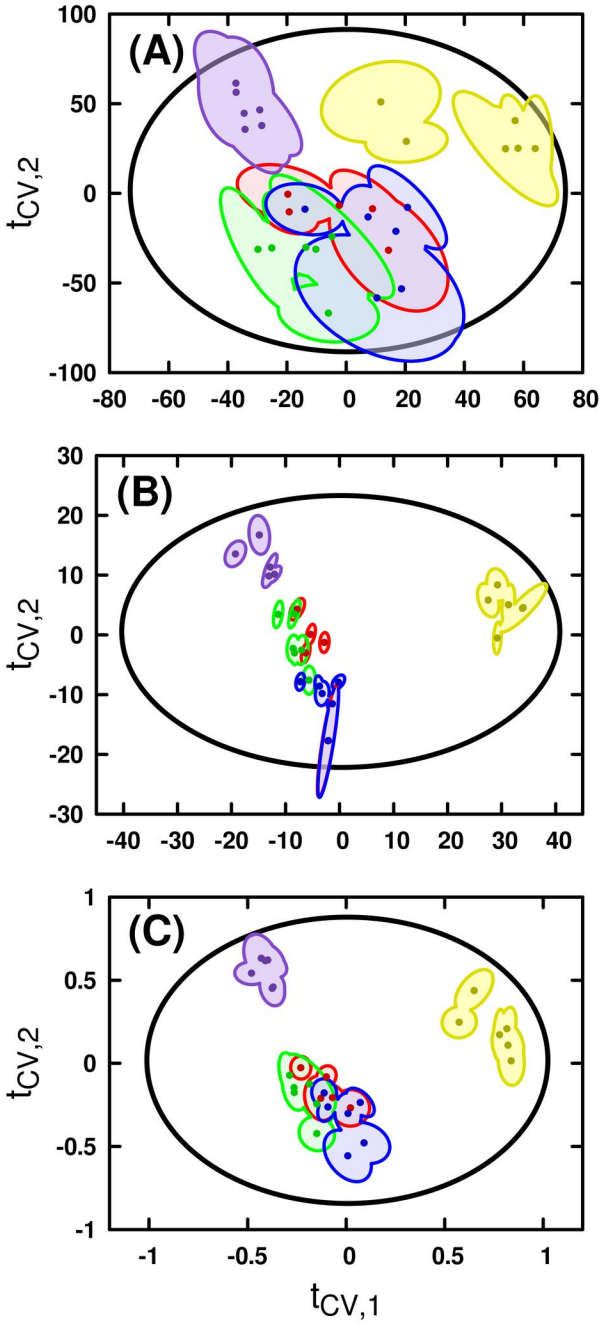

Figure 4.10: Comparison of PLS-DA and MB-PLS-DA Scores.
Cross-validated scores generated from (**A**) PLS-DA of ^{1}H NMR in vacuo, (**B**) PLS-DA of DI-ESI-MS in vacuo, and (**C**) MB-PLS-DA of ^{1}H NMR and DI-ESI-MS. Consensus directions in MB-PLS-DA scores space show decreased rotation during cross-validation when compared to PLS-DA scores of the in vacuo PCA model. Yellow, red, green, violet and blue scores correspond to the control, 6-OHDA, MPP^+, paraquat and rotenone classes, respectively.

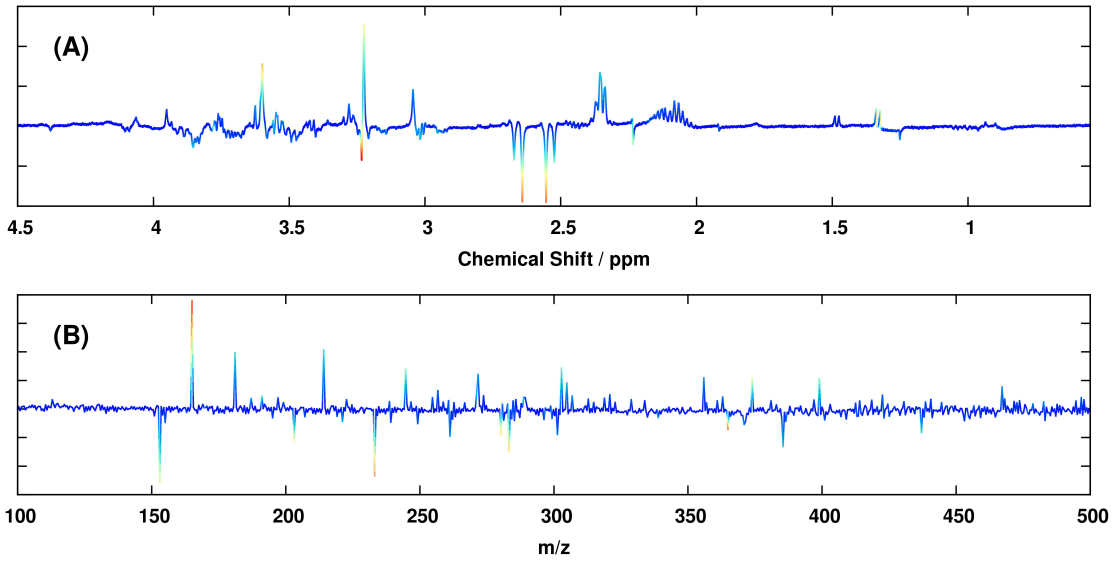

Figure 4.11: Backscaled NMR and MS Block Loadings.
Backscaled (**A**) ^{1}H NMR block and (**B**) DI-ESI-MS block loadings from MB-PLS-DA. Comparison of the above panels to those from MB-OPLS-DA (Figures 4.12, 4.13) reveals the mixed predictive and orthogonal variation present in MB-PLS loadings. It is also important to note that a second PLS component exists, and thus complete interpretation of the joint NMR and MS data requires simultaneous examination of *two* sets of block loadings.

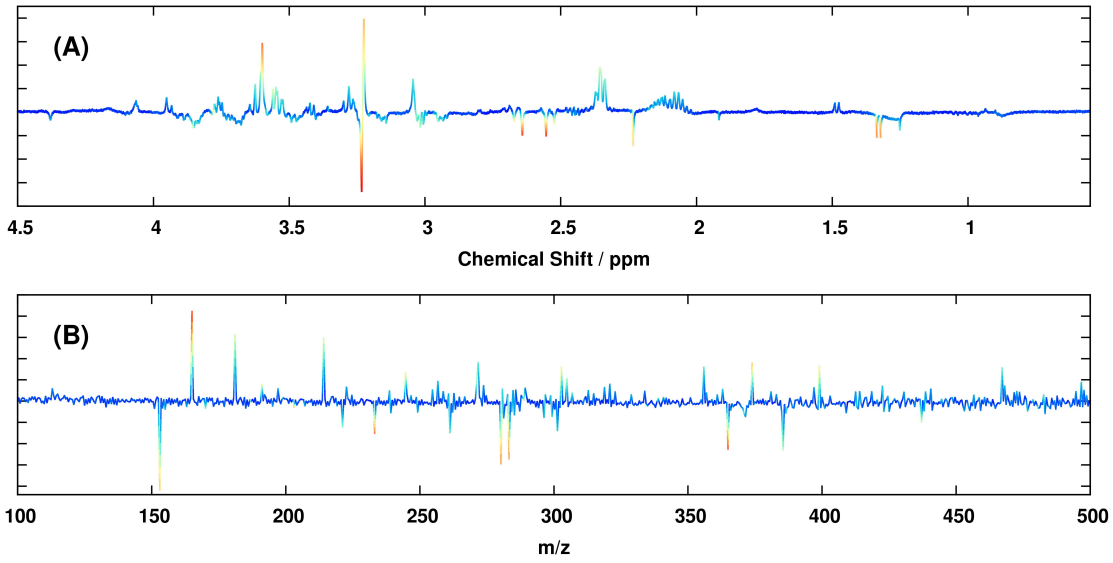

Figure 4.12: Backscaled NMR and MS Predictive Block Loadings.
Backscaled predictive (**A**) ^{1}H NMR block and (**B**) DI-ESI-MS block loadings from MB-OPLS-DA.

in the MB-OPLS model, providing richer information about the differing mechanisms of each neurotoxic drug. Additional orthogonal components would serve to further disentangle the two effects, at the slight expense of model reliability. Validation of the MB-OPLS model by CV-ANOVA resulted

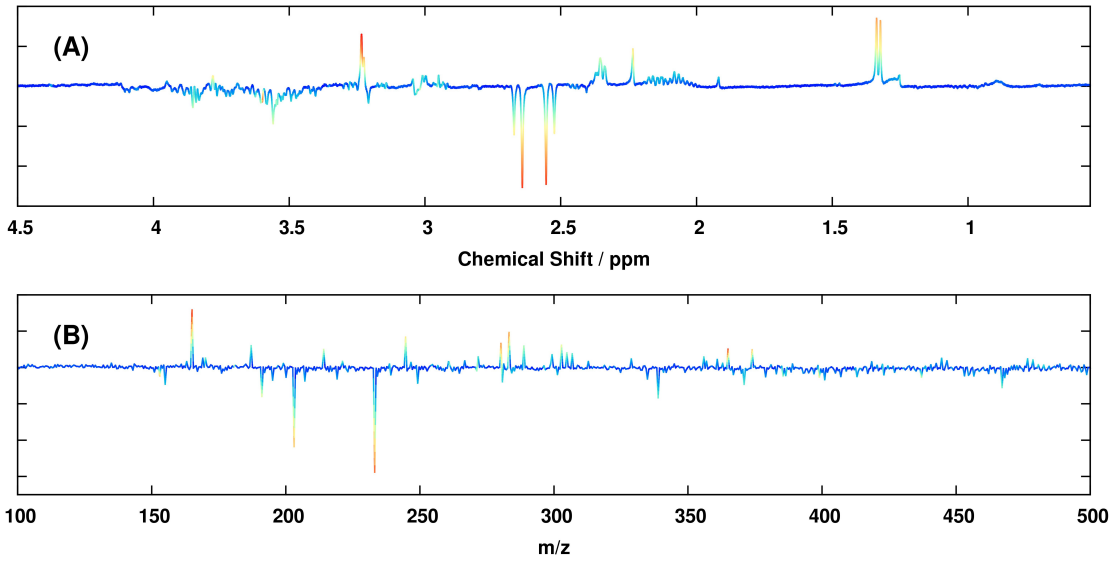

Figure 4.13: Backscaled NMR and MS Orthogonal Block Loadings.
Backscaled orthogonal (**A**) ^{1}H NMR block and (**B**) DI-ESI-MS block loadings from MB-OPLS-DA.

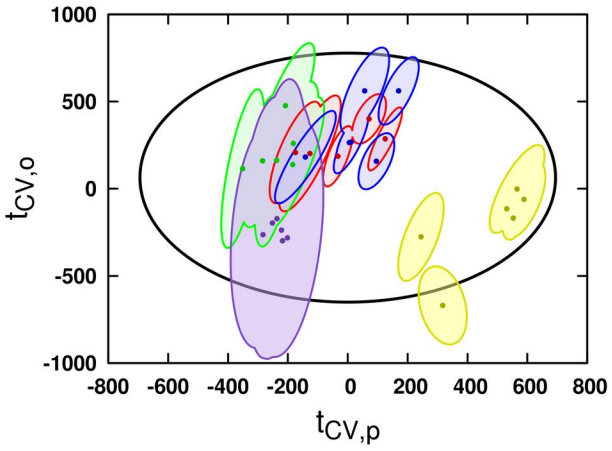

Figure 4.14: MB-OPLS-DA Cross-validated Scores.
Cross-validated scores generated from MB-OPLS-DA of the joint ^{1}H NMR and DI-ESI-MS data. The OPLS filter within MB-OPLS has effectively rotated the super-scores of the MB-PLS model (Figure 4.10C) to better differentiate between class-predictive and class-orthogonal variation.

in a p value equal to 5.5×10^{-6}, and permutation testing corroborated CV-ANOVA with $p < 0.001$, once again indicating a reliable supervised model.

4.3.3 Conclusions

The use of multiblock bilinear factorizations that capitalize on the availability of blocking information afforded greater model interpretability with the NMR and MS data than what was provided by single-block methods. The neurotoxins dataset provided an opportunity to compare the results of LOOCV and MCCV for optimal principal component count determination when marginally predictive data is being modeled. As expected, MCCV was a less optimistic estimator of model reliability than

LOOCV, and produced more parsimonious PCA decompositions. Finally, the dataset was an ideal proving ground for the new MB-OPLS algorithm, as MB-PLS-DA had clearly mixed class-predictive variation into multiple components. The use of MB-OPLS-DA resulted in more easily interpretable backscaled loadings, and provided more information relating to separations between control and drug treatment *and* separations between paraquat and other drug treatments (Figure 4.13).

4.4 Monte Carlo Analysis of Scores-space Separations

While the necessity of validating PLS and OPLS models is well understood within the statistics and chemometrics communities, it is an unfortunate fact that validation of PLS and OPLS models is still infrequent in work published by non-statistically oriented research groups [4]. This is especially true in the rapidly growing field of metabolomics, where these methods are quite often – and quite mistakenly – considered surrogates for PCA. PCA, PLS and OPLS are distinct modeling frameworks that achieve very different goals (cf. Section 3.5) and extract different information from a dataset. However, the optimistically forced class separations provided by PLS-DA and OPLS-DA have spawned a pattern of misuse in metabolomics and related fields. When PCA fails to identify significant separation between classes, untrained analysts may move to biased, insufficiently vetted OPLS-DA models without considering the statistical implications [2, 18]. While it is certainly possible for OPLS-DA to identify separation when PCA does not, the statistical significance of the separation must be validated before conclusions are drawn from the results. Studies that lack proper validation are automatically suspect from a statistical viewpoint, implying that future attempts to reproduce their results may fail. Thus, validation of all supervised models is an absolute requirement in chemometrics.

Even before supervised models are trained, the separations between classes in PCA scores space may be used as an informative qualitative predictor of whether reliable OPLS-DA models may be trained on the same data. This section presents practical guidelines on what level of OPLS-DA model reliability may be expected based solely on PCA class separations.

4.4.1 Materials and Methods

A Monte Carlo simulation was performed using MVAPACK [32] to analyze the relationship between class separations in PCA scores space and OPLS-DA cross-validation metrics, as a function of

spectral noise content. Two data matrices that both contained highly significant class-discriminating variation were used within two parallel simulations.

Initial Datasets

Two classes of observations (Light and Medium Decaffeinated) from the binned data matrix were extracted from the latest version of the Coffees dataset [32]. The resulting data matrix (referred to as $\mathbf{X}$: $N = 32, K = 284$) contains a highly significant separation between the two classes based on caffeine 1D ^{1}H NMR spectral features. A second dataset, generated from a comparison of two chemically defined cell growth media, was used to provide further support for the trends observed during Monte Carlo analysis of the Coffees data matrix. The resulting Media data matrix ($N = 50, K = 238$) also contains highly significant separation between two classes based on 1D ^{1}H NMR spectral features.

Prior to Monte Carlo simulation, the ℓ_2 norm (largest singular value) of each data matrix $\mathbf{X}$ was computed and stored as σ_{max}. A set of 50 noise standard deviations (σ) where each value ranged from $\sigma_{max}/500$ to $\sigma_{max}/10$. For each noise standard deviation, a set of 200 Monte Carlo iterations was performed. Another set of 200 iterations was also performed on each original data matrix $\mathbf{X}$ without any added noise.

Monte Carlo Simulation

At each Monte Carlo iteration, an $N \times K$ real matrix of noise values was drawn as NK independently and identically distributed samples from a zero-mean normal distribution having a standard deviation of σ, corresponding to the current noise value as described above. The data matrix $\mathbf{X}$ was summed with the noise matrix, and a three-component ($A = 3$) PCA model was computed on the resulting sum ($\mathbf{X}'$) after unit variance scaling [25] using a NIPALS algorithm [15]. The explained variation (R^2) of each principal component was computed as described in Section 3.6. A Monte Carlo leave-n-out cross-validation (MCCV) was performed based on the modified method of Krzanowski and Eastment [11] in order to obtain a per-component predictive ability (Q^2) statistic. A seven-fold partitioning of observations and variables, randomly resampled ten times, was performed for each PCA MCCV run. Following PCA model training, the Mahalanobis distance between the two classes was computed using PCA scores [6].

After computation of the Mahalanobis distance, the noisy data matrix $\mathbf{X}'$ was Pareto-scaled and subjected to OPLS-DA using a Pareto-scaled binary $(0, 1)$ response vector $(\mathbf{y})$ and a NIPALS OPLS algorithm [23]. A one-component $(A_p = 1, A_o = 1)$ OPLS model was constructed, from which backscaled predictive loadings were extracted by dividing by the coefficients obtained from Pareto scaling [5]. The Pearson correlation coefficient between backscaled loadings and the known "true" loadings – $\text{corr}(\mathbf{p}, \mathbf{p_0})$ – was computed for later visualization. Explained variation (R^2_Y) was computed as described in Section 3.6. A Monte Carlo leave-n-out internal cross-validation of the OPLS model was performed using a seven-fold partitioning of the data matrix that was randomly resampled ten times [34]. Predictive ability (DQ^2) statistics were computed as the mean DQ^2 obtained from MCCV results [29]. Thus, each OPLS model contained a set of ten fitted residual matrices from cross-validation available for use in CV-ANOVA significance testing [10]. During CV-ANOVA calculations, the median values of mean square error (MSE) were computed from all residual matrices, and the ratio of median fitted MSE to median residual MSE was calculated to yield an F-statistic for p value generation.

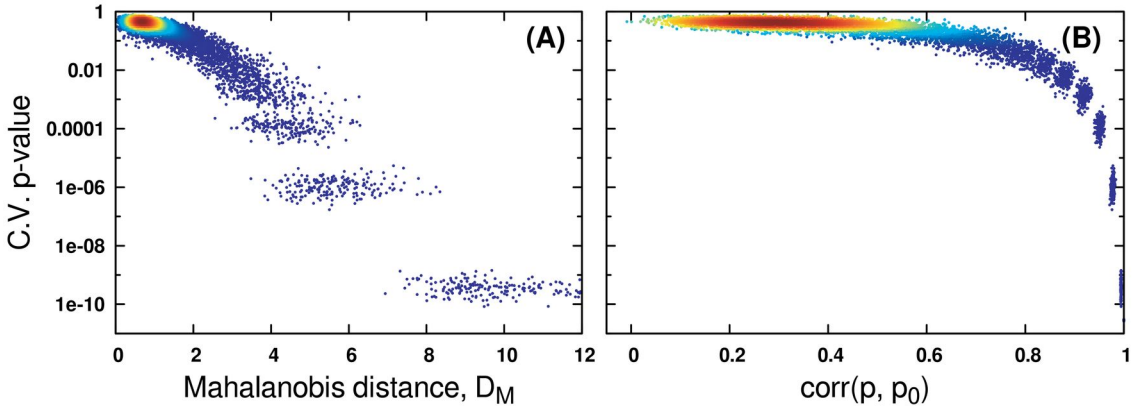

Figure 4.15: Monte Carlo Results for the Coffees Data Matrix.
Relationships to OPLS-DA CV-ANOVA p values obtained through Monte Carlo simulation of (**A**) the Mahalanobis distance (D_M) between classes in PCA scores space, and (**B**) the correlation between OPLS-DA model predictive loadings given noisy data $(\mathbf{p})$ and loadings obtained on the original Coffees data matrix $(\mathbf{p_0})$. The density of points in both panels is indicated by coloring, where red indicates high point density and blue indicates low density.

4.4.2 Results and Discussion

As expected, PCA scores-space class separations rapidly decreased as noise was added to the data. Addition of noise also forced a rise in OPLS-DA cross-validation statistics. As a result, a strong exponential relationship is observed between Mahalanobis distances calculated from PCA scores and

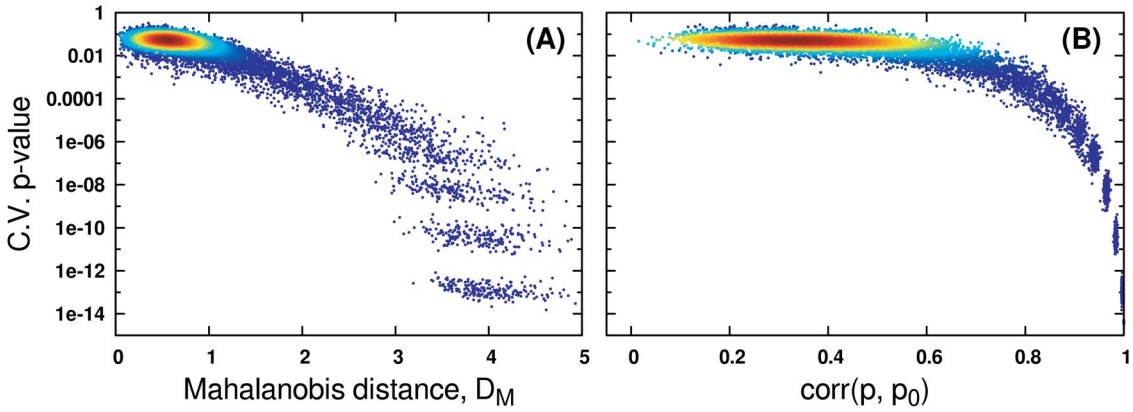

Figure 4.16: Monte Carlo Results for the Media Data Matrix.
Summary of Monte Carlo results from the Media data matrix. See Figure 4.15 for more details.

CV-ANOVA p values from OPLS-DA models (Figures 4.15A and 4.16A). Because PCA modeling uses no class membership information, the scores-space distances in these figures are essentially the least biased method of appraising discrimination ability. As the two classes become less distinguishable based on their spectral measurements, PCA will expose less separation between their scores. When PCA fails to expose class separation, OPLS-DA will continue to do so *at the expense of model reliability*, as it is relying on weaker sources of variation in the noisier data. While the exact form of the relationship between distance and p value will depend on the input data and responses, this analysis provides clear evidence that distances between classes in PCA scores may be used as a qualitative ruler of future supervised model reliability.

The shrinkage of Mahalanobis distances as data matrix noise increases occurs concomitantly with a rapid loss of correlation between ideal OPLS predictive loadings and estimated loadings (Figures 4.15B, 4.16B and 4.17A). It is critical to note that class separations in OPLS scores space do not appreciably decrease (Figure 4.18) with the decreased loading correlations. In effect, the OPLS model has identified different, *less reliable* sources of variation in the noisy data matrix in order to maintain class separation. OPLS-DA requires only that some variation in the measured data correlates with class membership, regardless of whether that variation is signal or noise [30, 23, 13]. When the true predictive spectral features that reflect the underlying biochemistry have become masked by noise, OPLS-DA will shift its focus to the variation that best predicts class membership. Because OPLS-DA provides the most optimistic result possible, validation becomes a necessity.

These Monte Carlo simulations once again illustrate how noise can masquerade as class-predictive variation in statistical analyses of high-dimensional spectral measurements. Moreover, the simula-

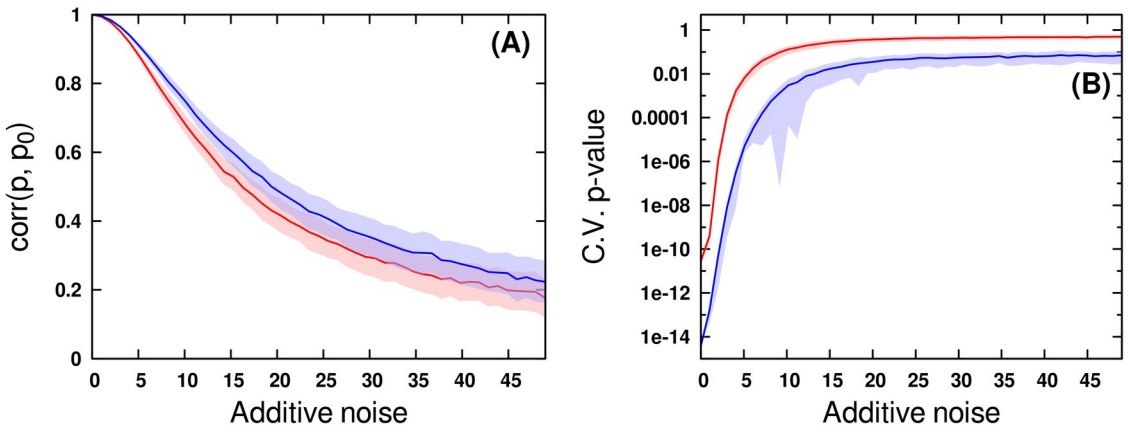

Figure 4.17: Effect of Noise on Loadings and CV-ANOVA Statistics.
(**A**) Decrease of correlation between estimated loadings (**p**) and true loadings ($\mathbf{p_0}$) as varying degrees of noise are added to the Coffees (red) and Media (blue) data matrices. Light shaded regions indicate confidence intervals of plus or minus one standard deviation from the mean correlation. A value of 1x additive noise corresponds to a noise standard deviation equaling 0.002 times the data matrix ℓ_2 norm. (**B**) Increase of p values from CV-ANOVA validation as varying degrees of noise are added to the data matrices. Shaded regions indicate plus or minus one standard deviation from the median p value.

tions touch on an often-overlooked distinction between class separations and *reliable, statistically significant* class separations in PCA/PLS scores space. Although PLS and OPLS may separate classes in situations where PCA cannot, this outcome should raise a red flag to the analyst that the model is suspect and the data may not sufficiently predict class membership. Only after rigorous cross-validation can it be safely inferred that OPLS-DA class separations are reliable and significant. If cross-validated estimates of OPLS-DA scores still separate the desired classes, and CV-ANOVA and permutation testing report significant p values, the models may be used for chemical inference. If cross-validation is left unreported, conclusions drawn from the models must be met with strong skepticism [27, 4].

The results of these Monte Carlo analyses relating PCA scores-space separations to OPLS-DA cross-validation metrics effectively summarize the reasons why rigorous cross-validation is necessary in chemometric studies relying on multivariate methods. More specifically, they reaffirm the importance of PCA as a first-pass unsupervised tool in metabolic fingerprinting and untargeted metabolic profiling studies, where class separations in scores space are often the sole basis for further experimentation. It is an unfortunate common practice in such studies to dismiss completely overlapped classes in PCA scores space and move ahead to (usually un-validated) supervised methods such as PLS and OPLS that force scores-space separation. Such practices almost guarantee the

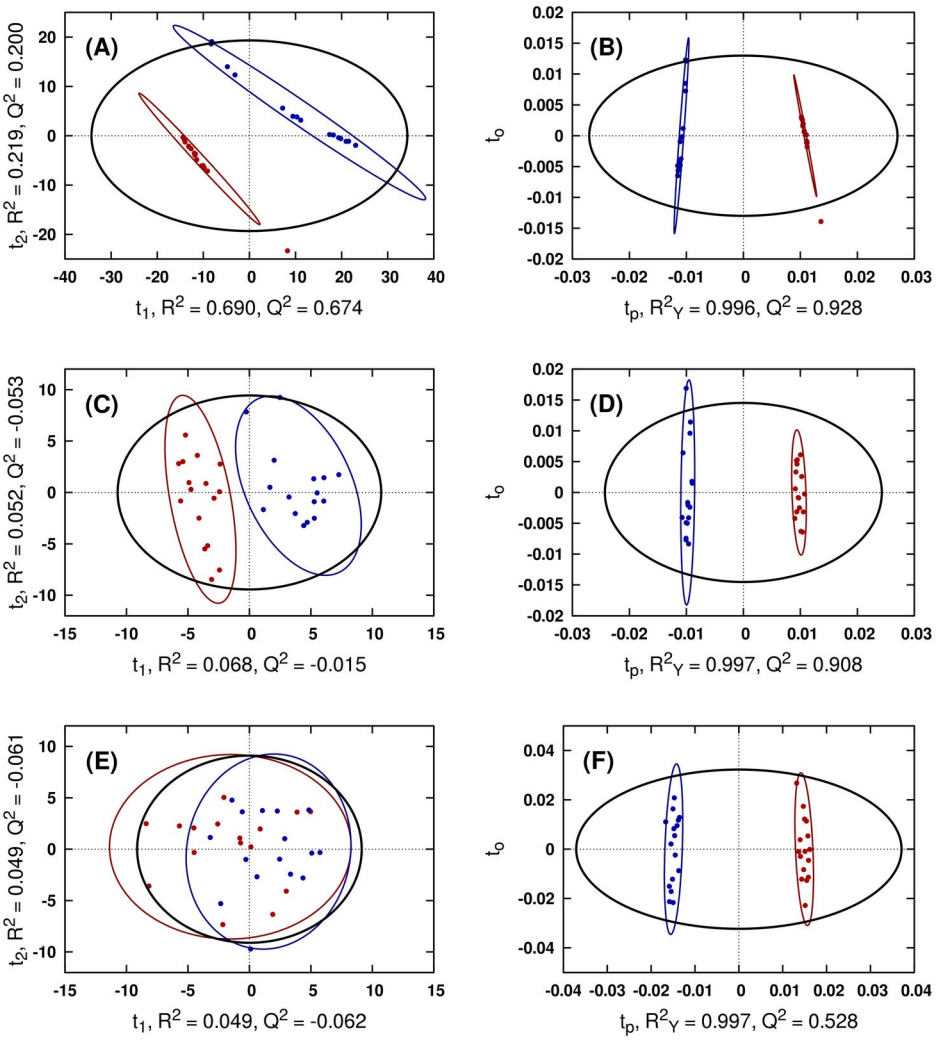

Figure 4.18: Effect of Noise on PCA and OPLS-DA Scores.
Comparison of representative PCA (**A**, **C**, **E**) and OPLS-DA (**B**, **D**, **F**) scores resulting from modeling the original data matrix (**A**, **B**), the 4x noisy data matrix (**C**, **D**), and the 20x noisy data matrix (**E**, **F**). Ellipses represent the 95% confidence regions for class membership.

irreproducibility of any conclusions drawn from trained multivariate models, as the relationship from Monte Carlo simulation indicates. It is therefore highly recommended that methods which assign Mahalanobis distance-based confidence ellipses to classes in PCA scores [31], report cross-validation estimated scores plots for PLS and OPLS models [27], and provide one or more cross-validated metrics during model training [32] be used in these studies whenever possible.

These analyses are only a case study for two specific data matrices, and are not meant to provide a quantitative relationship between any of the discussed metrics over all possible metabolomics stud-

ies. Instead, they lend positive numerical support to the recommendations that analysts rigorously validate their models by multiple means, including CV-ANOVA, response permutation testing, and even qualitative examination of PCA scores-space class separations. It is hoped that this work may be used to further promote best practices of supervised multivariate model training and validation in the community.

4.5 References

[1] K. M. Aberg, E. Alm, and R. J. Torgrip. The correspondence problem for metabonomics datasets. *Analytical and Bioanalytical Chemistry*, 394(1):151–162, 2009.

[2] A. A. Aksenov, L. Yeates, A. Pasamontes, C. Siebe, Y. Zrodnikov, J. Simmons, M. M. McCartney, J.-P. Deplanque, R. S. Wells, and C. E. Davis. Metabolite Content Profiling of Bottlenose Dolphin Exhaled Breath. *Analytical Chemistry*, 86(21):10616–10624, 2014.

[3] A. Belay, K. Ture, M. Redi, and A. Asfaw. Measurement of caffeine in coffee beans with UV/vis spectrometer. *Food Chemistry*, 108(1):310–315, 2008.

[4] R. G. Brereton. A short history of chemometrics: A personal view. *Journal of Chemometrics*, 28(10):725–736, 2014.

[5] O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, and J. Nicholson. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Analytical Chemistry*, 77(5):1282–1289, 2005.

[6] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[7] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[8] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

[9] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[10] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[11] P. Eshghi. Dimensionality choice in principal components analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems*, 130:6–13, 2014.

[12] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, 96(1):22–26, 2009.

[13] J. Gottfries, E. Johansson, and J. Trygg. On the impact of uncorrelated variation in regression mathematics. *Journal of Chemometrics*, 22:565–570, 2008.

[14] T. L. Hwang and A. J. Shaka. Water Suppression That Works – Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic Resonance*, 112(2):275–279, 1995.

[15] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[16] D. W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[17] D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015.

[18] G. McLaughlin, K. C. Doty, and I. K. Lednev. Raman Spectroscopy of Blood for Species Identification. *Analytical Chemistry*, 86(23):11628–11633, 2014.

[19] B. D. Nguyen, X. Meng, K. J. Donovan, and A. J. Shaka. SOGGY: Solvent-optimized double gradient spectroscopy for water suppression. A comparison with some existing techniques. *Journal of Magnetic Resonance*, 184(2):263–274, 2007.

[20] F. Rastrelli, S. Jha, and F. Mancin. Seeing through macromolecules: T_2-filtered NMR for the purity assay of functionalized nanosystems and the screening of biofluids. *Journal of the American Chemical Society*, 131(40):14222–14224, 2009.

[21] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[22] A. K. Smilde, J. A. Westerhuis, and S. de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337, 2003.

[23] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[24] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, W. R. Kent, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:402–408, 2008.

[25] R. a. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(142):1–15, 2006.

[26] J. A. Westerhuis and P. M. J. Coenegracht. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11(5):379–392, 1997.

[27] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[28] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[29] J. A. Westerhuis, E. J. J. van Velzen, H. C. J. Hoefsloot, and A. K. Smilde. Discriminant Q^2 (DQ^2) for improved discrimination in PLSDA models. *Metabolomics*, 4(4):293–296, 2008.

[30] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[31] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013.

[32] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[33] B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014.

[34] Q. S. Xu and Y. Z. Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

[35] Y. Xu, E. Correa, and R. Goodacre. Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection. *Analytical and Bioanaytical Chemistry*, 405(15):5063–5074, 2013.

[36] B. Zhang, S. Halouska, R. Gaupp, S. Lei, E. Snell, R. J. Fenton, R. G. Barletta, G. A. Somverville, and R. Powers. Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes. *Journal of Integrated OMICS*, 2(3):120–137, 2013.

[37] Y. Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.

Chapter 5
The MVAPACK Suite for NMR Chemometrics

5.1 Introduction

The biochemical laboratory procedures involved in metabolomics experiments are potentially straightforward and inexpensive, depending on the biological systems and pathways under study [49]. The minimal sample handling requirements of 1D ^{1}H NMR spectroscopy and the immense sensitivity of multivariate bilinear factorizations such as principal component analysis (PCA) and partial least squares (PLS) make NMR metabolic fingerprinting especially attainable. This low barrier to entry has no doubt contributed to the rapid recent growth of the field. Unfortunately, the data handling tasks of NMR metabolomics are far more difficult to properly execute. Commercial software packages available for multivariate analysis (e.g. SIMCA, PLS Toolbox, The Unscrambler, etc.) tend to be expensive and require more software for upstream processing and treatment of spectral data. Analysts are thus required to first open and process NMR data in packages such as ACD/1D NMR Manager (Advanced Chemistry Development), Mnova NMR (Mestrelabs Research) and perform further statistical treatment in MATLAB (The Mathworks, Natick, MA), R, or Microsoft Excel. This results in an unnecessarily cumbersome and time-consuming data handling pipeline by forcing the analyst to pass data between multiple software packages. As a result, the field of metabolomics research is littered with unpublished "in-house" software solutions created for processing, treating or modeling NMR datasets [38, 37, 7, 6, 11, 28, 43]. This continued reinvention of the wheel impedes progress in the field and complicates the tasks of standardization and communication of protocols that the metabolomics community is desperately attempting to achieve [29, 20]. Insult is then added to injury, as these in-house solutions are far less likely than their commercial counterparts to include proper means of validating trained multivariate models, further contributing to the general lack of model validation already present in the field [41]. While the community has released several official software packages targeted at metabolomics [26, 8, 39, 24, 46, 18, 1], none provide a complete, well-validated data path. At the time of this writing, no single software package existed to bring raw NMR data along its complete journey to validated, interpretable multivariate models.

These issues motivated the development of a free and open-source software package, MVAPACK, that provides a complete pipeline of functions for NMR chemometrics and metabolomics. MVAPACK is written in the GNU Octave mathematical programming language [14], which is also open-source and nearly syntactically identical to MATLAB. Thus, the installation of GNU/Linux, Octave and MVAPACK onto a commodity workstation provides a uniform environment in which a data analyst may truly work "from FIDs to models" in a few minutes using a set of well-documented, open-source, high-level data handling functions.

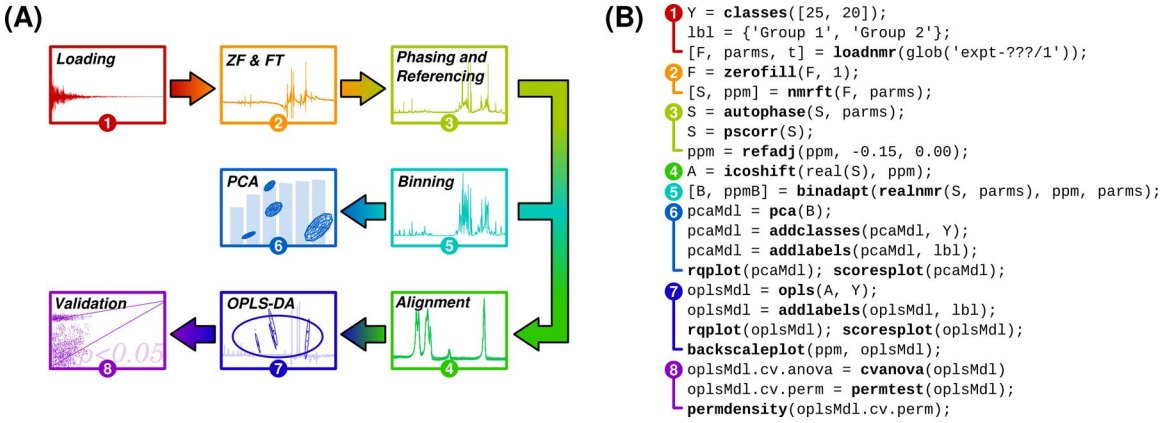

Figure 5.1: Example Data Handling Flow in MVAPACK.
A general NMR metabolic fingerprinting data handling flow diagram (**A**) and its associated minimum working example MVAPACK scripts (**B**). This minimalist data handling script is a simple starting point for using MVAPACK; much greater flexibility and functionality are present in the software than may be shown here. All functions in bold typeface are provided in MVAPACK.

5.2 Materials and Methods

5.2.1 Software Implementation

The MVAPACK software package is written in GNU Octave, an open-source mathematical programming language that uses MATLAB syntax [14]. Every function available in MVAPACK is realized as a single Octave function file that may be examined or changed using any text editor. Most functions in MVAPACK follow a similar input-to-output template, where an input data matrix **A** is modified and returned as an output data matrix **B**. Other input arguments, required or optional, may accompany **A**, and extra output values may accompany **B**, depending on the specific needs of the analyst. Models produced by PCA, PLS, OPLS, LDA, MB-PCA and MB-PLS are all similarly organized into Octave structures (i.e. "structs") that all follow scalar, vector, and matrix notations of Wold

et al. [45]. Thus, functions in MVAPACK are highly modular, often allowing drop-in replacement of one processing, treatment or modeling method for another by a simple change of function name and arguments. For instance, all modeling algorithms allow the specification of a scaling method at the time they are trained, so all available scaling functions conform to the same interface: for any input data matrix, a scaled data matrix is returned alongside the centering and scaling vectors used to compute the matrix.

Data may be handled in MVAPACK in either interactive mode, in which the user types commands into the Octave interpreter one at a time, or in batch mode, where a complete processing scheme has been laid out in an Octave script to be executed non-interactively. Once an ideal set of data handling steps is determined by interactive manipulation of any given dataset, it may be immortalized in an Octave script, thus providing documentation of procedures and allowing for rapid recalculation of all associated results.

Figure 5.1 illustrates a simple MVAPACK script capable of taking 1D ^{1}H NMR data from raw free induction decays to validated PCA and OPLS-DA models. In section 1, a binary class matrix **Y** and an accompanying set of class labels are built, and the time-domain raw data are loaded into the complex data matrix **F**. In section 2, the time-domain data matrix **F** is zero-filled once and Fourier-transformed to produce the complex spectral data matrix **S**. Section 3 automatically phase corrects each spectrum in **S**, normalizes and corrects for between-spectrum phase differences, and corrects the chemical shift abscissa to center the reference peak at 0.0 ppm. In sections 4 and 5, data handling splits into two pathways, where icoshift alignment [30] is used to generate a data matrix **A** fit for full-resolution OPLS-DA and AI-binning [9] is used to generate a data matrix **B** for PCA. In section 6, a PCA model is built and assigned classes and labels, and a model quality plot and a scores plot are produced. In section 7, similar functions are used to train an OPLS-DA model and produce summary plots. Finally, section 8 performs CV-ANOVA [15] and response permutation [41] significance tests to fully validate the supervised OPLS-DA model. While Figure 5.1 is completely functional, it is still an extremely bare-bones approach to metabolic fingerprinting. MVAPACK provides countless other functions and schemes for handling data.

5.2.2 Feature Set

The functions available in MVAPACK span the following general categories: data loading and processing (Table 5.1), treatment (Table 5.2), modeling (Table 5.3), and validation (Table 5.4) [20]. Specific features in each category are discussed in the following sections.

Table 5.1: MVAPACK Processing Feature Matrix.

	Topspin	VnmrJ	nmrPipe	NMRViewJ	MNova	ACD/NMR	Automics	Chenomx	KnowItAll	Metabonomic	MetaboAnalyst	AMIX	SIMCA	PLS Toolbox	PyChem	**MVAPACK**
Loading																
Bruker, 1D	*		*	*	*	*	*	*	*	*		*				*
Bruker, 2D	*		*	*	*	*										*
Varian, 1D		*	*	*	*	*	*	*	*			*				*
Apodization																
Exponential, 1D	*	*	*		*	*		*	*							*
Exponential, 2D	*	*	*		*	*										*
Gaussian, 1D	*	*	*		*	*			*							*
Gaussian, 2D	*	*	*		*	*										*
Sine, 1D	*	*	*		*	*										*
Sine, 2D	*	*	*		*	*										*
Zero-filling																
ZF, 1D	*	*	*		*	*	*	*	*							*
ZF, 2D	*	*	*		*	*										*
Transforms																
DFT, 1D	*	*	*		*	*	*	*	*	*						*
DFT, 2D	*	*	*		*	*										*
CWT, 1D					*								*			*
IST, 2D	*	*			*											*
Phase correction																
Manual, 1D	*	*	*		*	*	*	*	*	*		*				*
Manual, 2D	*	*	*		*	*						*				*
Automatic, 1D	*	*			*	*	*	*	*			*				*
Automatic, 2D	*	*			*	*										*

Processing

Loading of Bruker raw data is available using either a high-performance DMX-format loading routine or nmrPipe [10] as a backend, and loading of Varian data is available using an nmrPipe backend. Additionally, data in a variety of structured text formats may be read into MVAPACK using standard GNU Octave routines. The NMR spectral processing functions in MVAPACK follow the traditional paradigms of NMR processing [22] and include methods for apodization, zero-filling, Fourier

transformation, Iterative Soft Thresholding (IST) reconstruction [23], manual and automatic phase correction [31, 5, 2], region of interest selection and manipulation, peak picking [12], integration and referencing.

Table 5.2: MVAPACK Treatment Feature Matrix.

	Topspin	VnmrJ	nmrPipe	NMRViewJ	MNova	ACD/NMR	Automics	Chenomx	KnowItAll	Metabonomic	MetaboAnalyst	AMIX	SIMCA	PLS Toolbox	PyChem	**MVAPACK**
Binning																
Uniform, 1D					*	*	*		*	*		*				*
Uniform, 2D						*	*		*			*				*
Optimized, 1D												*				*
Adaptive, 1D												*				*
Adaptive, 2D																*
Alignment																
Global							*									*
Interval							*									*
Normalization																
CS											*	*		*	*	*
PQ											*					*
HM																*
SNV							*		*	*		*		*	*	*
MSC							*		*					*	*	*
PSC																*
Scaling																
UV			*				*		*	*	*	*	*	*	*	*
Pareto					*		*		*	*	*		*			*
Range					*				*	*	*		*			*
Level					*								*			*
VAST					*				*				*			*

Treatment

Functions for statistical data treatment in MVAPACK include binning [33, 9], alignment [30], normalization [3, 11, 34], scaling [36], and direct orthogonal signal correction [40]. In addition, MVAPACK supports uniform binning and AI-binning (Chapter 6) of third-order data tensors stored as arrays of real matrices in Octave.

Table 5.3: MVAPACK Modeling Feature Matrix.

	Topspin	Vnmr.J	nmrPipe	NMRViewJ	MNova	ACD/NMR	Automics	Chenomx	KnowItAll	Metabonomic	MetaboAnalyst	AMIX	SIMCA	PLS Toolbox	PyChem	**MVAPACK**
Bilinear																
PCA			*		*		*		*	*	*	*	*	*	*	*
LDA							*					*				*
PLS							*			*	*	*	*	*		*
OPLS													*	*		*
Multiblock																
MB-PCA														*		*
MB-PLS														*		*
MB-OPLS																*

Modeling

MVAPACK provides complete support for building PCA [27], LDA [21], PLS [44, 19, 45], OPLS [35, 4], MB-PCA and MB-PLS [42, 32] models from processed and treated datasets. At the current time, only bilinear factorizations are supported within MVAPACK.

Table 5.4: MVAPACK Validation Feature Matrix.

	Topspin	Vnmr.J	nmrPipe	NMRViewJ	MNova	ACD/NMR	Automics	Chenomx	KnowItAll	Metabonomic	MetaboAnalyst	AMIX	SIMCA	PLS Toolbox	PyChem	**MVAPACK**
Validation																
R^2, Q^2							*		*	*	*	*	*	*	*	*
Permutation											*		*	*		*
CV-ANOVA													*			*
Visualization																
2D Scores					*		*		*	*	*	*	*	*	*	*
3D Scores									*	*	*		*	*		*
Loadings					*		*		*	*	*	*	*	*	*	*
Backscaling																*
S-plot													*			*
SUS-plot													*			*
Data Ellipsoid										*	*	*	*	*		*
Class Ellipsoid																*

Validation

All PCA and MB-PCA models are validated as they are built based on the results of a Monte Carlo leave-n-out (Modified $K + E$) internal cross-validation [13, 16], and all PLS, OPLS and MB-PLS models are validated during training based on results of a Monte Carlo leave-n-out internal cross-validation [47, 48]. In all cases, a set of R^2 (i.e. R^2_X and R^2_Y) statistics are generated to assess how well each data and response matrix is approximated by the models, and Q^2 statistics are generated to describe self-consistency and predictive ability of each data matrix in PCA and PLS models, respectively. Per-component R^2 and Q^2 statistics are utilized by MVAPACK to estimate the optimal number of model components. Because cross-validation is performed using a Monte Carlo scheme in MVAPACK, all Q^2 statistics are reported with confidence intervals, regardless of model type. Further validation of supervised models is available in the form of CV-ANOVA [15] and response permutation [41] significance testing, both of which report p values that indicate model validity.

5.3 Discussion and Conclusions

This chapter presents MVAPACK, a completely free and open-source data handling environment tailor-suited to NMR chemometrics and ^{1}H NMR and MS metabolic fingerprinting applications. Unlike data handling chains composed of multiple commercial software packages, MVAPACK is free to use, modify and distribute according to the GNU General Public License [17] and provides a single consistent data handling environment. Because MVAPACK is written for GNU Octave, researchers already familiar with MATLAB syntax will also be familiar with MVAPACK without a considerable learning curve. Datasets and results obtained using MVAPACK are readily saved and exchanged using GNU Octave built-in support for the MATLAB *mat*-file format.

A recent review [25] of software packages targeted at metabolomics highlights the piecemeal nature of 1D ^{1}H NMR data handling in the field, where no single package is capable of performing all the tasks required by the analyst. MVAPACK addresses this need by providing a complete pipeline that is tuned for metabolic fingerprinting. Use of MVAPACK reduces data analysis time in metabolic fingerprinting from days to minutes, simply by collecting all the required functions into a single scriptable environment. In fact, the example script in Figure 5.1 would execute in under five minutes on a modern GNU/Linux or Mac OS X computer system.

The routine processing of *any* 1D and 2D NMR spectral data may be readily done with MVAPACK, and processing routines are easily batched. The MVAPACK scripts written to analyze the datasets in Chapter 4 are composed of intuitive, modular commands that logically subdivide the script into recognizable tasks like automatic phase correction, spectral alignment, normalization, and so forth. Furthermore, aside from physical memory limitations of the host computer, MVAPACK does not impose any limit on the number of NMR observations that may be simultaneously handled.

A major advantage of MVAPACK is the seamless transfer of the processed, treated NMR data to multivariate statistical analyses. The PCA, PLS, OPLS and LDA bilinear modeling algorithms, now ubiquitous in the metabolomics community, are all implemented in MVAPACK using a consistent under-the-hood framework. Model results may be visualized and interpreted using MVAPACK routines that provide scatter, line and bar plots of model scores, loadings and validation statistics. Critically, MVAPACK automatically ensures that *all* trained models are valid using leave-one-out and Monte Carlo leave-n-out internal cross-validation routines and provides further means of validating supervised models in the form of CV-ANOVA and response permutation significance testing. Several powerful examples of MVAPACK applied to real datasets are presented in Chapter 4. Because it implements well-established algorithms available from peer-reviewed chemometrics literature, MVAPACK generates identical results when compared to expensive software packages like Umetrics SIMCA-P+.

In short, MVAPACK provides a complete platform for NMR chemometric data handling that is ideal for both routine handling of metabolomics datasets and development of novel algorithms. Unlike its closed-source predecessors, the modular, open-source design of MVAPACK readily accepts new functionality, allowing it to grow and maintain pace with the state-of-the-art in the chemometrics literature. MVAPACK is freely available for download at `http://bionmr.unl.edu/mvapack.php`. Detailed documentation of MVAPACK, all datasets presented in Chapter 4, and the scripts used to handle them are also available for download.

5.4 References

[1] A. Alonso, M. A. Rodriguez, M. Vinaixa, R. Tortosa, X. Correig, A. Julia, and S. Marsal. Focus: A robust workflow for one-dimensional NMR spectral analysis. *Analytical Chemistry*, 86(2):1160–1169, 2014.

[2] G. Balacco and C. Cobas. Automatic phase correction of 2D NMR spectra by a whitening method. *Magnetic Resonance in Chemistry*, 47(4):322–327, 2009.

[3] R. J. Barnes, M. S. Dhanoa, and S. J. Lister. Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, 43(5):772–777, 1989.

[4] M. Bylesjo, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20(8-10):341–351, 2006.

[5] L. Chen, Z. Q. Weng, L. Y. Goh, and M. Garland. An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1-2):164–168, 2002.

[6] O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, and J. Nicholson. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Analytical Chemistry*, 77(5):1282–1289, 2005.

[7] O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson, and E. Holmes. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. *Analytical Chemistry*, 77(2):517–526, 2005.

[8] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, and B. Walczak. TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemometrics and Intelligent Laboratory Systems*, 85(2):269–277, 2007.

[9] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[10] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe – a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular NMR*, 6(3):277–293, 1995.

[11] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

[12] P. Du, W. a. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006.

[13] H. T. Eastment and W. J. Krzanowski. Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*, 24(1):73–77, 1982.

[14] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[15] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[16] P. Eshghi. Dimensionality choice in principal components analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems*, 130:6–13, 2014.

[17] F. S. Foundation. GNU General Public License, version 3. `http://www.gnu.org/licenses/gpl.html`, June 2007. Last retrieved 2015-05-11.

[18] E. Gaude, F. Chignola, D. Spiliotopoulos, A. Spitaleri, M. Ghitti, J. M. Garcia-Manteiga, S. Mari, and G. Musco. muma, An R Package for Metabolomics Univariate and Multivariate Statistical Analysis. *Current Metabolomics*, 1(2):180–189, 2013.

[19] P. Geladi and B. R. Kowalski. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[20] R. Goodacre, D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjostrom, J. Trygg, and F. Wulfert. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241, 2007.

[21] W. Hardle and L. Simar. *Applied multivariate statistical analysis*. Springer, 2012.

[22] J. C. Hoch and A. S. Stern. *NMR Data Processing*. Wiley, 1996.

[23] S. G. Hyberts, G. J. Heffron, N. G. Tarragona, K. Solanky, K. A. Edmonds, H. Luithardt, J. Fejzo, M. Chorev, H. Aktas, K. Colson, K. H. Falchuk, J. A. Halperin, and G. Wagner. Ultrahigh-Resolution 1H-13C HSQC Spectra of Metabolite Mixtures Using Nonlinear Sampling and Forward Maximum Entropy Reconstruction. *Journal of American Chemical Society*, 129(16):5108–5116, 2007.

[24] J. L. Izquierdo-Garcia, I. Rodriguez, A. Kyriazis, P. Villa, P. Barreiro, M. Desco, and J. Ruiz-Cabello. A novel R-package graphic user interface for the analysis of metabonomic profiles. *BMC Bioinformatics*, 10, 2009.

[25] J. L. Izquierdo-Garcia, P. Villa, A. Kyriazis, L. del Puerto-Nevado, S. Perez-Rial, I. Rodriguez, N. Hernandez, and J. Ruiz-Cabello. Descriptive review of current NMR-based metabolomic data analysis packages. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 59(3):263–270, 2011.

[26] R. M. Jarvis, D. Broadhurst, H. Johnson, N. M. O'Boyle, and R. Goodacre. PYCHEM: A multivariate analysis package for python. *Bioinformatics*, 22(20):2565–2566, 2006.

[27] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[28] J. Kang, M. Choi, S. Kang, H. Kwon, H. Wen, C. H. Lee, M. Park, S. Wiklund, H. J. Kim, S. W. Kwon, and S. Park. Application of a ^{1}H Nuclear Magnetic Resonance (NMR) Metabolomics Approach Combined with Orthogonal Projections to Latent Structure-Discriminant Analysis as an Efficient Tool for Discriminating between Korean and Chinese Herbal Medicines. *Journal of Agricultural and Food Chemistry*, 56(24):11589–11595, 2008.

[29] J. C. Lindon, J. K. Nicholson, E. Holmes, H. C. Keun, A. Craig, J. T. M. Pearce, S. J. Bruce, N. Hardy, S. A. Sansone, H. Antti, P. Jonsson, C. Daykin, M. Navarange, R. D. Beger, E. R. Verheij, A. Amberg, D. Baunsgaard, G. H. Cantor, L. Lehman-McKeeman, M. Earll, S. Wold, E. Johansson, J. N. Haselden, K. Kramer, C. Thomas, J. Lindberg, I. Schuppe-Koistinen, I. D. Wilson, M. D. Reily, D. G. Robertson, H. Senn, A. Krotzky, S. Kochhar, J. Powell, F. van der Ouderaa, R. Plumb, H. Schaefer, and M. Spraul. Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology*, 23(7):833–838, 2005.

[30] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[31] M. M. Siegel. The Use of the Modified Simplex-Method for Automatic Phase Correction in Fourier-Transform Nuclear Magnetic-Resonance Spectroscopy. *Analytica Chimica Acta*, 5(1):103–108, 1981.

[32] A. K. Smilde, J. A. Westerhuis, and S. de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337, 2003.

[33] S. A. A. Sousa, A. Magalhaes, and M. M. C. Ferreira. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122:93–102, 2013.

[34] R. J. O. Torgrip, K. M. Aberg, E. Alm, I. Schuppe-Koistinen, and J. Lindberg. A note on normalization of biofluid 1D ^{1}H NMR data. *Metabolomics*, 4(2):114–121, 2008.

[35] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[36] R. a. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(142):1–15, 2006.

[37] K. C. M. Verhoeckx, S. Bijlsma, S. Jespersen, R. Ramaker, E. R. Verheij, R. F. Witkamp, J. dan Der Greef, and R. J. T. Rodenburg. Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis. *International Immunopharmacology*, 4(12):1499–1514, 2004.

[38] M. R. Viant. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, 310(3):943–948, 2003.

[39] T. Wang, K. Shao, Q. Y. Chu, Y. F. Ren, Y. M. Mu, L. J. Qu, J. He, C. W. Jin, and B. Xia. Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics*, 10, 2009.

[40] J. A. Westerhuis, S. de Jong, and A. K. Smilde. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 56(1):13–25, 2001.

[41] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[42] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[43] S. Wiklund, E. Johansson, L. Sjostrom, E. J. Mellerowicz, U. Edlund, J. P. Shockcor, J. Gottfries, T. Moritz, and J. Trygg. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, 80(1):115–122, 2008.

[44] S. Wold, E. Johansson, and M. Cocchi. *PLS: Partial Least Squares Projections to Latent Structures*. KLUWER ESCOM Science Publisher, 1993.

[45] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[46] J. G. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, and D. S. Wishart. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 40(W1):127–133, 2012.

[47] Q. S. Xu and Y. Z. Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

[48] Q. S. Xu, Y. Z. Liang, and Y. P. Du. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18(2):112–120, 2004.

[49] B. Zhang, S. Halouska, R. Gaupp, S. Lei, E. Snell, R. J. Fenton, R. G. Barletta, G. A. Somverville, and R. Powers. Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes. *Journal of Integrated OMICS*, 2(3):120–137, 2013.

Chapter 6
Phase-Scatter Correction of NMR Datasets

6.1 Introduction

As previously introduced in Chapter 3, normalization of data tensors is a commonly performed procedure aimed at minimizing the within-class variation of two or more groups of observations, relative to the total or between-class variation in the dataset. Irrespective of whether separations between classes are obtained using an unsupervised PCA model or a supervised (O)PLS-DA model, greater statistical significance and increased biological relevance may be ascribed to separations between classes having greater variation between groups than within them [11].

Normalization applied directly to hypercomplex NMR data (or its real component) is sub-optimal, as even small phase differences between observations can frustrate the estimation of normalization factors (cf. Section 3.3). Possibly worse, blind normalization of poorly phased spectral data can accentuate experimentally irrelevant spectral features in a data tensor during multivariate modeling, leading the analyst to erroneous conclusions. These difficulties motivated the development of phase-scatter correction (PSC, [13]) as a means of simultaneously correcting the coupled phase errors and dilution errors that are present in hypercomplex NMR data tensors. When hypercomplex NMR data must be normalized prior to multivariate analyses within the confines of a metabolomics study, the interrelation of phase and dilution errors is best handled using phase-scatter correction.

6.2 Theory

6.2.1 Multiplicative Scatter Correction

Phase-scatter correction (PSC) is effectively an extension of multiplicative scatter correction (MSC) to handle phase errors during normalization. In MSC, each real spectrum is scaled around its mean intensity and shifted to match a reference spectrum, typically the mean of the dataset [3]. Optimal normalization factors ($\mathbf{b}$) of a data matrix $\mathbf{X}$ are determined by a linear regression of the

mean-centered reference vector onto the mean-centered matrix:

$$\big(\mathbf{X} - \langle\mathbf{X}\rangle\big)^T \mathbf{b} = \big(\mathbf{r} - \langle\mathbf{r}\rangle\big)^T \tag{6.1}$$

where observations are stored as row vectors in $\mathbf{X}$, and $\mathbf{r}$ is the reference observation row vector. The equation above represents an overdetermined system of linear equations, therefore the least-squares estimate of $\mathbf{b}$ may be computed rapidly, and MSC is rather computationally efficient.

6.2.2 Phase-scatter Correction

PSC additionally corrects zero- and first-order phase errors during normalization, requiring a non-linear optimization of the following objective:

$$Q(\mathbf{X} \mid \mathbf{p}) = \sum_{n=1}^{N} \|\mathbf{z}_n \circ \mathbf{x}_n - \mathbf{r}\|_2^2 \tag{6.2}$$

where $\circ$ denotes the element-wise product, the mean-centered matrix $\mathbf{X}$ lies in $\mathbb{H}_1^{N\times K}$, the mean-centered reference $\mathbf{r}$ lies in $\mathbb{H}_1^K$, and the set of parameters $\mathbf{p}$ includes a normalization factor and two phase errors per observation in $\mathbf{X}$:

$$\mathbf{p} = \{b_1, \ldots b_N, \theta_{0,1}, \ldots \theta_{0,N}, \theta_{1,1}, \ldots \theta_{1,N}\} \tag{6.3}$$

and each vector $\mathbf{z}_n$ contains the normalization and phase corrections for the n-th observation $\mathbf{x}_n$:

$$z_{n,k} = b_n e^{u_1(\theta_{0,n} + \theta_{1,n} k)} \tag{6.4}$$

Because the reference observation $\mathbf{r}$ is fixed during optimization, minimization of $Q(\mathbf{X} \mid \mathbf{p})$ may be achieved by independently minimizing each observation's contributions. Minimization is carried out for every observation in the data matrix using Levenberg-Marquardt nonlinear least squares [7] as implemented by the *leasqr* function in GNU Octave, a function similar to MATLAB's *nlinfit*. Each corrected spectrum $\hat{\mathbf{x}}_n$ is then returned from optimization as follows:

$$\hat{\mathbf{x}}_n = \mathbf{z}_n \circ \mathbf{x}_n + \langle\mathbf{r}\rangle \tag{6.5}$$

Phase-scatter correction of 50 1D ^{1}H NMR spectra having 32,768 complex points each requires

approximately 30 seconds on a single-core 3.2 GHz Intel workstation running GNU Octave 3.6.

6.2.3 Ensemble Phase Correction

It is important recognize that the phase-scatter correction objective function $Q(\mathbf{X} \mid \mathbf{p})$ provides no measure of ideal phase values, meaning that PSC requires an additional phase correction step prior to its execution in order to ensure adequate initial conditions. Even when $\mathbf{X}$ has been suitably phase-corrected, PSC may still attempt to minimize scatter between spectra by re-introducing phase errors. This undesirable behavior of PSC may be observed when large disparities in spectral intensities are present between observations. To correct this, a standard phase correction objective $f : \mathbb{H}_D^K \to \mathbb{R}$ may be combined with the PSC objective using a Lagrange multiplier, like so:

$$\Lambda(\mathbf{X} \mid \mathbf{p}) = -\sum_{n=1}^{N} f(\boldsymbol{\theta}_n \circ \mathbf{x}_n) + \lambda \sum_{n=1}^{N} \|\mathbf{z}_n \circ \mathbf{x}_n - \langle \mathbf{Z} \circ \mathbf{X} \rangle\|_2^2 \tag{6.6}$$

where the correction matrix $\mathbf{Z}$ has the same form as in PSC, expressed as a real diagonal matrix of normalization factors $\mathbf{B}$ and a hypercomplex matrix of phase factors $\boldsymbol{\Theta}$:

$$\mathbf{Z} = \mathbf{B}\boldsymbol{\Theta} \tag{6.7}$$

and $\boldsymbol{\theta}_n$ is the n-th row of $\boldsymbol{\Theta}$. The new ensemble phase correction (EPC) objective function $\Lambda(\mathbf{X} \mid \mathbf{p})$ balances the potentially opposing goals of phase correction and scatter correction through the Lagrange multiplier λ, and does not require the specification of a reference observation $\mathbf{r}$. In effect, EPC allows its scatter correction reference to float as the current mean of the data, $\langle \mathbf{Z} \circ \mathbf{X} \rangle$. This floating reference requires the simultaneous optimization of all the parameters in $\mathbf{p}$, unlike phase-scatter correction. Efficient minimization of $\Lambda(\mathbf{X} \mid \mathbf{p})$ may be accomplished by a modified Nelder-Mead simplex optimization procedure [8], which serially updates the simplices of all observations at each global iteration and maintains the current mean vector $\langle \mathbf{Z} \circ \mathbf{X} \rangle$ at each update.

In contrast to phase-scatter correction, which seeks to minimize the scatter of data matrix observations around a fixed reference, ensemble phase correction approaches the dilemma of entwined phase and normalization errors from an opposing direction by introducing a scatter term into a standard automatic phase correction procedure. The amount of normalization achieved by EPC is directly controlled by the magnitude of λ: in the opposite limits of $\lambda = 0$ and $\lambda \to \infty$, EPC be-

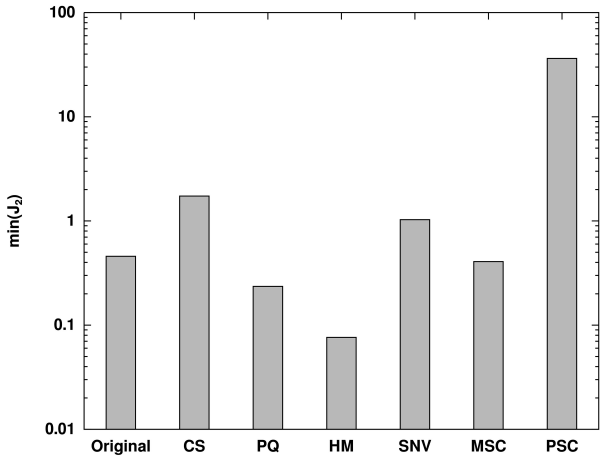

Figure 6.1: Cluster Quality after Normalization and PCA Modeling. Comparison of PCA cluster quality for ^{1}H NMR metabolomics data normalized using different algorithms. The minimum J_2 value (worst cluster quality) for each model is reported here, as it is a more effective indicator of overall model and cluster quality than the mean or median.

comes equivalent to standard phase correction and phase-scatter correction with a floating reference, respectively.

6.3 Materials and Methods

6.3.1 NMR Data Processing

Previously collected ^{1}H NMR spectral data from published work [4] was leveraged as a typical metabolomics dataset for performance analysis of PSC versus other normalization methods. Free induction decays were loaded into GNU Octave 3.6 [2] for processing using MVAPACK routines [12]. Time-domain signals were zero-filled to 32,768 points and Fourier transformed, resulting in a complex data matrix of 177 spectra divided among 16 classes (N = 177, K = 32,768, M = 16). Spectra were both automatically phase corrected by simplex entropy minimization [1] and manually phase corrected by applying a constant phase shift to all spectra. Both automatically and manually phase corrected datasets were then normalized using the CS, PQ, HM, SNV, MSC and PSC methods (cf. Chapter 3). Each normalized data matrix was binned using a uniform 0.04 ppm bin width, scaled per-variable to unit variance, and subjected to PCA. The J_2 statistic [6] was calculated for each class to provide a measure of cluster quality for the PCA scores from each normalization method, as follows:

$$J_{2,m} = \frac{|\mathbf{C}|}{|\mathbf{C}_m|} \tag{6.8}$$

where $\mathbf{C}_m$ is the covariance matrix of the scores in class m, $\mathbf{C}$ is the covariance matrix of all scores, and the vertical bars represent the matrix determinant. Thus, as a cluster shrinks relative to the entirety of the scores-space data, its J_2 statistic will increase. While J_2 provides a measure

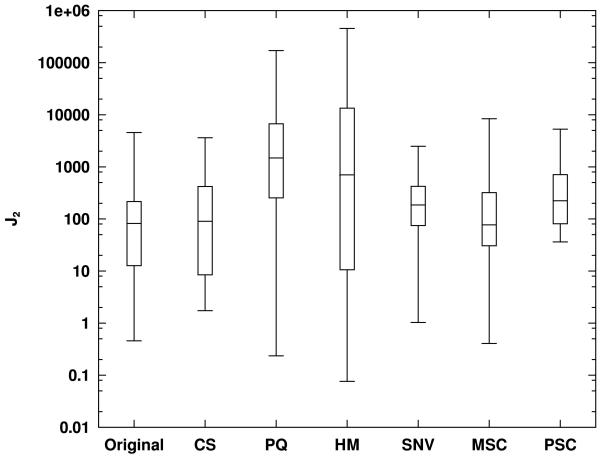

Figure 6.2: Cluster Quality after Normalization and PCA Modeling. Comparison of PCA cluster quality for ^{1}H NMR metabolomics data normalized using different algorithms. For each normalization method, a box is defined by the estimated first and third quartiles of J_2 for the clusters and whiskers are defined by the range of the J_2 values for the clusters. For this dataset, the minimum J_2 value is most instructive, given the fact that overall model quality is not well-reflected by the J_2 metric in the case of distorted principal components.

of individual cluster tightness, it does not capture the degree of cluster overlap within a dataset. Figures 6.1 and 6.2 show the results of the J_2 calculation for normalization methods applied to real ^{1}H NMR metabolomics data.

To quantify differences between extracted principal components of automatically and manually phase corrected datasets, the angle between the first principal component loading vector of each pair of models (φ) was calculated as follows:

$$\varphi = \cos^{-1}\left(\mathbf{p}_{auto}{}^{T}\mathbf{p}_{man}\right) \tag{6.9}$$

where $\mathbf{p}_{auto}$ and $\mathbf{p}_{man}$ are the first-component loadings computed from a given normalization method's data after automatic and manual phase correction, respectively. The loading angle φ for a given normalization method is a reflection on that method's ability to properly normalize data and produce consistent PCA models from different initial phase error conditions.

Table 6.1: Metabolite Spectra Used in Monte Carlo Simulations.

Aminobutyrate	Adenosine	Alanine	Arginine
Asparagine	Aspartate	Choline	Citrulline
Ethanolamine	Fructose	Galactose	Glucose
Glutamate	Glutamine	Glycine	Histidine
Isoleucine	Lactate	Leucine	Lysine
Malate	Maltose	Myoinositol	Ornithine
Phenylalanine	Proline	Putrescine	Serine
Succinate	Sucrose	Threonine	Valine

6.3.2 Simulated NMR Datasets

The ^{1}H NMR spectra of 100 mM samples of 32 metabolites (Table 6.1) at pH 7.4 were downloaded from the Biological Magnetic Resonance Bank (BMRB, [10]) and fit to mixtures of complex Lorentzian functions using ACD/1D NMR Processing (Advanced Chemistry Development). Peak amplitudes (A), chemical shifts (ω_0), and widths (λ) returned from fitting were loaded into GNU Octave to generate simulated spectra having 65,536 data points and a spectral width of 11 ppm, centered at 4.7 ppm, based on the following model function:

$$s(\omega_k) = \sum_{p=1}^{P} \frac{A_p \lambda_p}{\lambda_p + u_1(\omega_k - \omega_{0,p})} \tag{6.10}$$

where $s(\omega_k)$ is the k-th data point of the spectrum, P equals the number of peaks, and u_1 equals the imaginary unit. Spectra were referenced and normalized to the DSS peak, and peaks corresponding to HOD and DSS were subsequently removed, resulting in a basis set of 32 perfectly-phased, noise-free metabolite spectra. Finally, the basis metabolite spectra were stored row-wise in a matrix $\mathbf{S}$ for later use in Monte Carlo calculations.

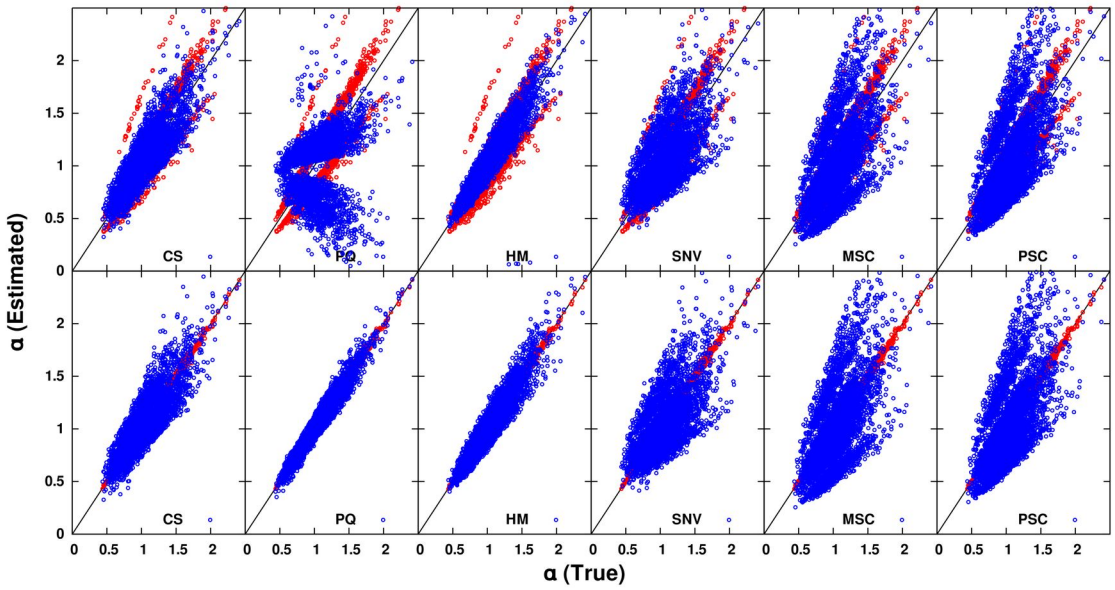

Figure 6.3: Monte Carlo Normalization Results.
Results of 100 Monte Carlo iterations at 0.2° zero-order phase error, indicating the ability of all normalization methods to recover the true dilution factor of a nearly perfectly phased dataset. Red points reflect the dilution factors calculated by integrated the DSS peak and blue points reflect the dilution factor estimates from normalization. Upper panels show the dilution factors recovered from automatically phased data after normalization, and lower panels show dilution factors recovered from unphased data after normalization.

6.3.3 Monte Carlo Experiments

Using the basis metabolite spectra, a dataset of 48 simulated metabolomics spectra ($\mathbf{X} \in \mathbb{H}_1^{N\times K}$) was generated according to the following equation:

$$\mathbf{X} = \mathbf{A}\left(\mathbf{CS} + \mathbf{1r}^T\right) + \mathbf{E} \tag{6.11}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a diagonal matrix of dilution factors α_n, $\mathbf{C} \in \mathbb{R}^{N\times P}$ is a matrix of metabolite concentrations, $\mathbf{S} \in \mathbb{H}_1^{P\times K}$ is the previously created metabolite basis set, $\mathbf{r} \in \mathbb{H}_1^{K}$ is a spectrum of the DSS reference peak, $\mathbf{1} \in \mathbb{R}^{N}$ is a vector of ones, and $\mathbf{E} \in \mathbb{H}_1^{N\times K}$ is a matrix of complex Gaussian white noise. Dilution factors were drawn from a log-normal distribution having zero mean and $\sigma = 0.25$. Concentrations in $\mathbf{C}$ were drawn from normal distributions with parameters chosen to mimic those in Torgrip et al. (Table 6.2) [9]. The resultant data in $\mathbf{X}$ is a simulated set of $N = 48$ metabolite extracts, spiked with 100 μM DSS, where six distinct classes arise from differences in the concentrations of alanine, asparagine, glutamate, malate, proline, sucrose and valine. All other metabolites were assigned concentrations from a normal distribution having $\mu = 5$ μM and $\sigma = 0.5$ μM.

Table 6.2: Metabolite Concentrations Altered in Monte Carlo Simulations.

Metabolite	$\mathbf{C_A}$ (μM)	$\mathbf{C_B}$ (μM)	$\mathbf{C_C}$ (μM)	$\mathbf{C_D}$ (μM)	$\mathbf{C_E}$ (μM)	$\mathbf{C_F}$ (μM)
Alanine	9.2 ± 1.4	19.6 ± 1.6	16.9 ± 1.2	6.5 ± 0.66	26.2 ± 3.6	13.5 ± 1.1
Asparagine	6.8 ± 0.86	11.7 ± 1.8	19.0 ± 1.9	14.7 ± 1.2	24.8 ± 2.6	17.4 ± 1.0
Glutamate	13.3 ± 1.7	9.2 ± 1.5	18.8 ± 1.9	16.9 ± 2.1	25.0 ± 3.5	6.9 ± 1.0
Malate	14.2 ± 1.2	11.9 ± 1.4	22.0 ± 5.1	6.7 ± 0.68	9.4 ± 0.72	18.0 ± 2.4
Proline	11.4 ± 1.5	18.4 ± 3.1	14.7 ± 2.4	6.9 ± 0.62	9.8 ± 1.5	23.7 ± 2.9
Sucrose	7.1 ± 0.9	17.2 ± 2.1	19.3 ± 2.0	13.2 ± 1.9	9.3 ± 0.56	23.3 ± 2.7
Valine	9.0 ± 0.85	26.3 ± 2.3	13.4 ± 1.2	20.4 ± 1.7	6.7 ± 0.90	17.0 ± 1.5

Monte Carlo simulations were run to assess the performance of all discussed normalization methods over various amounts of phase error added to $\mathbf{X}$. Forty-six phase error points were calculated, in which the standard deviation of θ_0 was linearly increased from 0° to 5°. The standard deviation of θ_1 at each point was equal to one tenth that of θ_0. Both θ_0 and θ_1 were assigned zero mean. For each phase error point, 100 Monte Carlo iterations were performed with different sets of random dilution factors. Spectra in the de-phased $\mathbf{X}$ matrix were automatically phase corrected using simplex entropy minimization and normalized each time using CS, PQ, HM, SNV, MSC and PSC methods. Normalization to unit DSS integral was also performed for reference. An identical set of

normalization calculations was performed on the unphased data. Estimated dilution factors were compared to the true values to produce a root-mean-square dilution error, $RMSE(\alpha)$, for each method. Figure 6.3 shows the $RMSE(\alpha)$ result of Monte Carlo simulation at 0.2° phase error. To assess normalization effects on multivariate model quality, spectra from each method were uniformly binned with 0.04 ppm bin widths, each bin scaled to unit variance, and subjected to PCA. Values of J_2 for each of the six classes were then calculated, and the median of the values was reported for each Monte Carlo iteration. The φ values between automatically phased and unphased principal component loadings were also calculated at each iteration to asses each normalization method's ability to produce consistent models in the presence of phase errors. Figure 6.4 summarizes the results of Monte Carlo simulation over all phase errors based on $RMSE(\alpha)$, J_2 and φ.

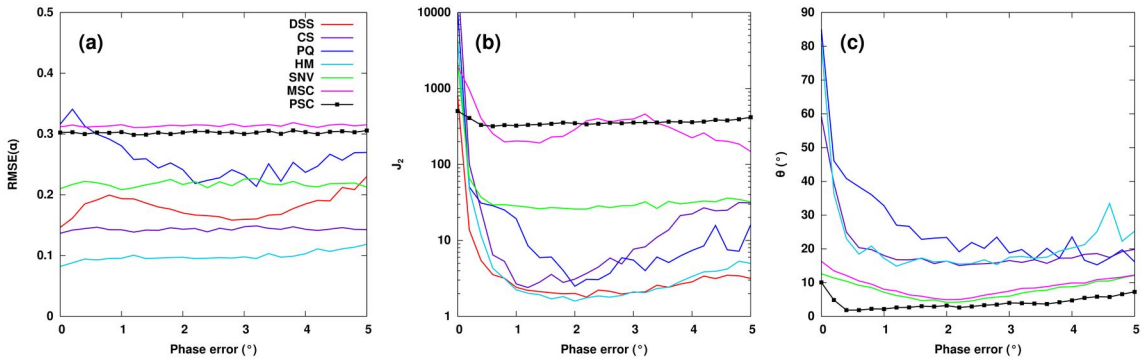

Figure 6.4: Summary of Monte Carlo Simulation Results.
Results of the Monte Carlo simulation over all phase error points. (**A**) As phase error increases, dilution factor estimates from all methods except PQ remain fairly stable. Estimates from PSC compete with MSC, but suffer in comparison with HM. (**B**) However, J_2 values indicate that PSC outperforms all other normalization methods at producing tight clusters at any realistic phase error. (**C**) Finally, values of φ calculated from PCA loadings indicate that PSC maintains the highest model consistency in the face of imperfectly phased data. Phase error on the x-axis refers to zero-order error; it should be noted that each point also contains first-order phase error as discussed in the Methods.

6.4 Results

On the real metabolomics spectral data, PSC normalization resulted in the highest quality clusters (Figure 6.5) according to the J_2 statistic shown in Figure 6.1. Given the fact that the spectra were each automatically phase corrected before any normalization was applied, this observed increase in J_2 must be due to the correction of subtle phase differences *between* spectra not detectable by correcting each spectrum individually. It is important to note that, while PQ and HM produce higher median J_2 values (Figure 6.2), this is an artifact of large distortions of their respective PCA loadings, and not always reflective of higher-quality clusters. Because J_2 is a per-cluster statistic,

it is only an ideal measure of overall scores-space model quality when all clusters are nearly identically distributed. Models containing highly distorted components may contain several high-quality clusters and a few extremely low-quality clusters, resulting in a high mean or median J_2 value. For that reason, the lower bound of J_2 for each method – effectively the worst cluster quality – was chosen as a better indicator of overall model quality than the median. In fact, PSC produced the most consistent model loadings between automatically and manually phase corrected data, with a φ value of 14.5°. This can be compared to φ values of 89.6° and 20.2° for PQ and HM, respectively.

Moreover, Monte Carlo analysis of PSC versus contemporary normalization methods show that PSC offers a unique advantage during multivariate analysis. Results of Monte Carlo normalization after automatic phase correction and summarized in Figure 6.4, and scatter plots of recovered dilution factors are shown in Figure 6.3. While PSC fails to recover true dilution factors as accurately as DSS, CS or HM normalization, it does remain competitive with MSC at all phase errors (Figure 6.4A). PSC normalization yields tighter clusters than all other methods, as is apparent from Figure 6.4B. Furthermore, PSC results in dramatically lower values of φ than all other methods, indicating that residual phase errors left uncorrected by automatic phase correction are significant enough to distort principal component loadings when normalized by any method other than PSC (Figure 6.4C).

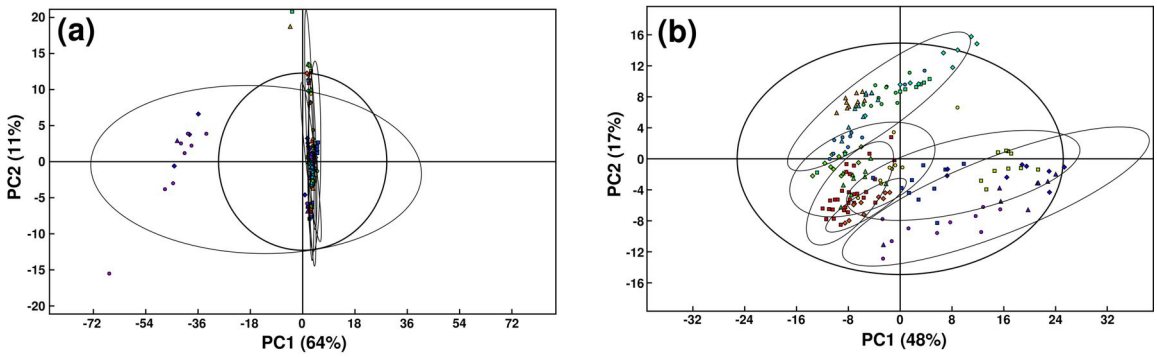

Figure 6.5: Distortion of Principal Components by PQ Normalization.
PCA scores of a typical metabolomics dataset after automatic phasing followed by either PQ or PSC normalization. In both plots, ellipses denote different classes of antibiotic treatment of *Mycobacterium smegmatis* and differing symbols within each ellipse represent different antibiotic subclasses. (**A**) PQ normalization amplifies residual phase differences left behind after automatic phasing, but (**B**) PSC normalization produces a more interpretable PCA model by correcting residual phase differences.

6.5 Discussion

As is evident from visual inspection of both the real metabolomics dataset and the Monte Carlo simulated datasets, correction of minute phase differences between spectra yields a substantial improvement in cluster quality in multivariate analyses. In general, phase differences contribute significantly to spectral line shape difference in ^{1}H NMR data. This effect is especially pronounced in the case of PSC normalization of spectra containing significant and consistent broad background signals, where normalization alone cannot comparably standardize baselines (cf. Chapter 7).

One particularly striking result of the Monte Carlo simulations is the difference between automatically phase corrected and unphased dilution factor estimates (Figure 6.3). In fact, examination of dilution factors estimated by DSS integration clearly shows that automatic phase correction introduces variation into the dataset through minute differences in θ_0 and θ_1 between spectra. This artificial variation is then amplified through normalization, as is especially apparent in the case of PQ normalization.

In their report on HM normalization, Torgrip et al. noticed the potential unsuitability of the explained sum of squares (R^2_X) for assessing model quality differences due to normalization methods [9]. As a ratio measure, explained sum of squares is not suitable for comparing the qualities of PCA models trained on different data, or any preprocessing done prior to building the models [5]. Therefore, the J_2 statistic was chosen as an alternative means of comparing cluster quality during Monte Carlo simulation. Effectively, J_2 measures the ratio of the area of a cluster in scores space relative to the total scores-space area, regardless of how much variation the model captures. Even still, because J_2 is a per-cluster statistic, it is not an ideal measure of overall scores-space model quality, especially for models containing highly distorted components. Mean or median J_2 values of a model may be high in this case, despite the fact that the model scores are useless from the perspective of class discrimination. Thus, the minimum J_2 was chosen as a more effective indicator of overall cluster quality.

6.6 Conclusions

Phase-scatter correction is a novel algorithm for simultaneously correcting zero- and first-order phase errors and random dilution factors in ^{1}H NMR chemometric data. While PSC only performs

comparably to MSC in dilution factor estimation, it more consistently yields high-quality clusters and interpretable models when used prior to PCA decomposition. PSC can be fully automated through prior automatic phase correction of the dataset, has no tunable parameters, and makes no assumptions regarding line shape, baseline flatness, or intensity distributions in the data. These qualities lend PSC to use in chemometrics as a new method of normalizing NMR data entering into multivariate analyses such as PCA or PLS. The latest implementation of PSC is available in the MVAPACK toolbox [12].

6.7 References

[1] L. Chen, Z. Q. Weng, L. Y. Goh, and M. Garland. An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1-2):164–168, 2002.

[2] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[3] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, 96(1):22–26, 2009.

[4] S. Halouska, R. J. Fenton, R. G. Barletta, and R. Powers. Predicting the in vivo Mechanism of Action for Drug Leads Using NMR Metabolomics. *ACS Chemical Biology*, 7(1):166–171, 2012.

[5] K. Kjeldahl and R. Bro. Some common misunderstandings in chemometrics. *Journal of Chemometrics*, 24(7-8):558–564, 2010.

[6] K. Koutroumbas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2006.

[7] D. W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[8] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, pages 308–313, 1964.

[9] R. J. O. Torgrip, K. M. Aberg, E. Alm, I. Schuppe-Koistinen, and J. Lindberg. A note on normalization of biofluid 1D ^{1}H NMR data. *Metabolomics*, 4(2):114–121, 2008.

[10] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, W. R. Kent, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:402–408, 2008.

[11] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013.

[12] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[13] B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014.

Chapter 7
Uncomplicated Statistical ^{1}H NMR Spectral Remodeling

7.1 Introduction

Structure-activity relationships (SAR) by NMR [22] spurred a revolution for the role of NMR in drug discovery. Like X-ray crystallography, NMR had been primarily used as a means to determine protein and protein-ligand structures as part of a structure-based drug discovery effort [4]. NMR is now an important alternative to traditional high-throughput screening (HTS) assays for identifying drug-like chemical leads [16, 17]. By combining NMR ligand-affinity screens with fragment-based libraries, a dramatic increase in chemical diversity is achieved (from 10^6 to 10^{63}), while also minimizing resources, increasing hit-rates and improving the drug-like qualities of the resulting chemical leads [5]. Consequently, NMR fragment-based screens have significantly benefited the pharmaceutical industry by leading to a number of clinical-stage compounds.

NMR ligand-affinity screening is also a powerful platform for protein functional annotation during the search for novel drug targets [13, 20]. Significant percentages of the human proteome and the proteomes of other infectious organisms are comprised of functionally uncharacterized proteins [14]. Undoubtedly hidden among this multitude of unannotated proteins are novel drug targets that may lead to new treatments or new means of overcoming mechanisms of drug resistance. Besides verifying that a functionally unannotated protein is druggable, NMR ligand affinity screening also identifies the functional epitope and the classes of ligands that bind the uncharacterized protein. This information may then be leveraged to infer a function through structural similarities with functionally annotated proteins [18, 19].

NMR spectroscopy reports a multitude of time-averaged physical observables that carry information relating to the nature of interactions between small molecule ligands and protein targets [10]. A number of 1D ^{1}H NMR pulse sequences have been developed to probe these distinct features of binding, including differences in free and bound ligand diffusion and relaxation properties [6], and

saturation transfers from water [1] and protein [11] resonances. As part of an NMR high-throughput screen, these 1D ^{1}H NMR pulse sequences present a number of unique challenges that include high false positive rates, long acquisition times, and high demand for protein samples [9, 7]. However, at suitably chosen concentrations of ligand and protein, a standard, unedited 1D ^{1}H NMR experiment may be used to detect binding interactions through enhanced relaxation rates of ligand spins [12, 20, 13].

While it is possible to detect ligand binding using standard 1D ^{1}H NMR, the resulting spectra are a combination of free and bound ligand and protein signals, a fact which makes them difficult to interpret. Broad, rolling baselines arising from slowly tumbling protein spins are particularly problematic during interpretation, as they often mask changes in ligand signal broadness and intensity. This masking effect due to protein baselines is exacerbated at protein-ligand concentration ratios nearing or exceeding unity, forcing the use of excess ligand and increasing the false negative rate during screening. To mitigate these issues, a statistical method called Uncomplicated Statistical Spectral Remodeling (USSR), was developed that removes protein baselines from high-throughput ligand-based screening datasets by leveraging inter-sample reproducibility of protein signals. In addition, it will be demonstrated that the use of phase-scatter correction greatly improves inter-sample protein baseline reproducibility and reduces the false-positive rate incurred by subsequent USSR-based analyses. The combination of PSC and USSR enables a rapid analysis of standard 1D ^{1}H NMR screening data, especially in difficult cases having a high protein-ligand concentration ratio.

7.2 Materials and Methods

7.2.1 Sample Preparation and NMR Acquisition

A set of 117 samples containing free ligand mixtures and a set of 117 samples containing Bovine Serum Albumin (BSA) with ligand mixtures were prepared based on previously published procedures [20, 13]. In summary, each mixture contained no more than four ligands, each ligand had a concentration of 100 μM, and BSA had a concentration of 200 μM when present. All NMR samples were prepared to 600 μL total volume in a buffer containing 10 mM bis-tris-d_{19}, 1.0 mM NaCl, 1.0 mM KCl, 1.0 mM $MgCl_2$ and 10 μM trimethylsilyl propanoic acid (TMSP) in D_2O at pH 7.0 (uncorrected). Samples were loaded into standard 5 mm NMR tubes for spectral acquisition.

All NMR experiments were collected on a Bruker Avance DRX 500 MHz spectrometer equipped

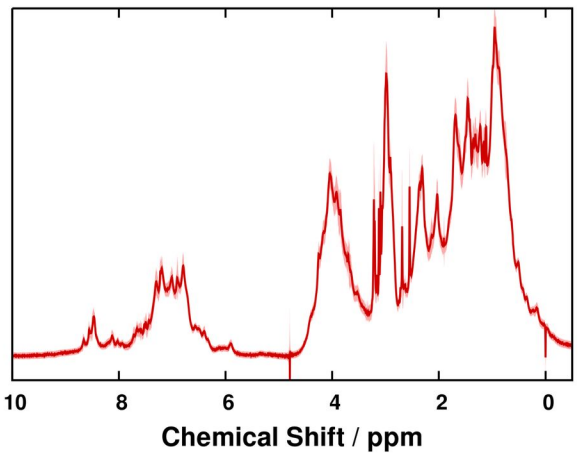

Figure 7.1: Statistical Baseline from the BSA Screening Dataset.
Statistical baseline ($\boldsymbol{\mu} \pm 4\boldsymbol{\sigma}$) computed from the ^{1}H NMR ligand-based screen against BSA. The mean baseline is traced in deep red, while the confidence region for the baseline is filled in light red underneath.

with a 5 mm inverse triple-resonance (^{1}H, ^{13}C, ^{15}N) cryoprobe with a z-axis gradient. A Bruker BACS-120 sample changer and ICON-NMR software were used to automate NMR data collection. Standard 1D ^{1}H NMR spectra were collected for each sample using a SOGGY water suppression pulse sequence [8, 15]. All experiments were performed at 20°C with 256 scans, 8 dummy scans, a carrier frequency offset of 2,352 Hz, a 5,483 Hz spectral width, and a 1.0 section inter-scan delay. Free induction decays were collected with 4,096 complex data points, resulting in a total acquisition time of 8 minutes per experiment.

7.2.2 NMR Data Processing

Acquired NMR spectra were loaded and processed in batch inside the GNU Octave 3.6 programming environment [3] using functions available in the MVAPACK software package [23]. Free induction decays were loaded in from Bruker DMX binary format and corrected for group delay by a fixed circular shift. All decays were then zero-filled twice, Fourier transformed and automatically phase corrected using a simplex optimization routine. Phase-scatter correction was applied to a copy of the screen spectral data, and spectral remodeling was performed in parallel on the uncorrected and corrected datasets for the purposes of comparison.

7.2.3 Statistical Spectral Remodeling

The Uncomplicated Statistical Spectral Remodeling (USSR) method capitalizes on the reproducibility of the protein baseline and the low likelihood that ligand signals will dominate any given spectral data point across multiple samples. For each pair of free mixture ($\mathbf{f}_n$) and screen (mixture plus protein, $\mathbf{p}_n$) ^{1}H NMR spectra, a difference spectrum ($\mathbf{d}_n$) was computed using a simple point-wise

subtraction. The central tendency ($\boldsymbol{\mu}$) and dispersion ($\boldsymbol{\sigma}$) of the difference spectra were then robustly estimated using the median and median absolute deviation, respectively. Figure 7.1 shows the statistical baseline computed by USSR from a screen of ligand binding to BSA. Once a statistical baseline is established for a given dataset, each spectrum $\mathbf{p}_n$ in the screen is remodeled to maximally remove interference from baseline signals. Each spectral data point in $\mathbf{p}_n$ is compared to $\boldsymbol{\mu} \pm \boldsymbol{\sigma}$ using a Bonferroni-corrected Student's t-test [2]. The resulting p value provides a measure of how distinguishable the corresponding data point is from the statistical baseline. Based on a preselected level of significance (α), data points having low p values are retained (less the statistical baseline) in the remodeled spectrum ($\mathbf{r}_n$) and data points having high p values are modeled as Gaussian white noise. Figure 7.2 shows an example remodeled spectrum from the ligand binding analysis of BSA.

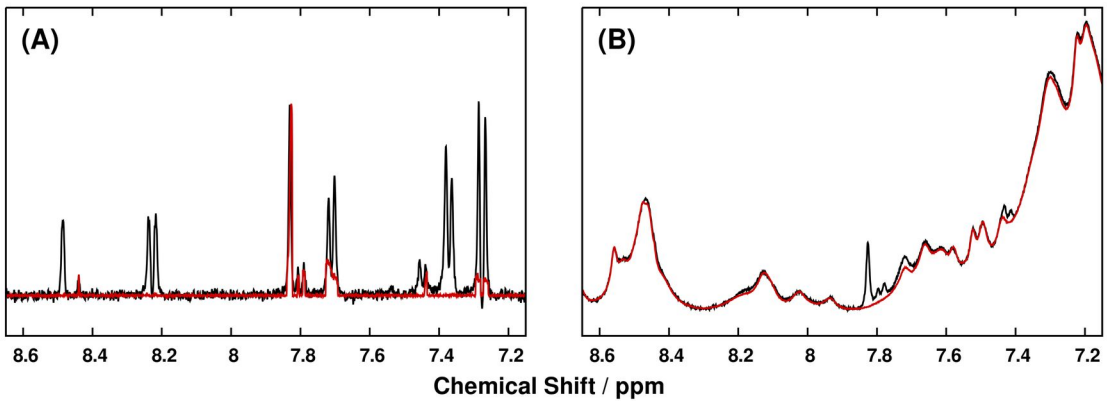

Figure 7.2: Statistical Baseline Removal from a Screen Spectrum.
An example spectral remodeling result of tolazamide, dimethyl 4-methoxyisophthalate, 1,7-dimethylxanthine and oxolinic acid in the presence of BSA, showing (**A**) the free ligand spectrum (black) and the remodeled spectrum (red) resulting from removing the statistical baseline (red) from the screen spectrum (black) in (**B**). The remodeled pseudospectrum readily indicates that several peaks from dimethyl 4-methoxyisophthalate have broadened into the baseline due to interaction with BSA.

7.2.4 Statistical Hit Determination

For each peak in each remodeled spectrum from USSR, a K_D was computed based on the intensity ratio between free and remodeled ligand signals. First, in the limit of fast exchange between free and bound ligand states relative to the NMR timescale, the fraction of bound ligand (f_B) was computed:

$$f_B = \left(\frac{I_F}{I_B} - 1\right)\left(\frac{v_B}{v_F} - 1\right)^{-1} \tag{7.1}$$

where I_F and I_B are the intensities of free and remodeled (bound) ligand signals, and v_F and v_B are the estimated NMR line widths of the free and remodeled ligand signals, respectively [21]. This fast-exchange assumption may be safely regarded as valid in most high-throughput 1D ^{1}H NMR protein-ligand affinity screening experiments [10], where the width and intensity of each ligand signal is a population-weighted sum of its values in the free and bound states. Without any assumptions about relative concentrations of ligand and protein, the fraction of bound ligand is related to the total protein concentration $[P]_T$, total ligand concentration $[L]_T$ and K_D via the following equation [21]:

$$f_B = \left[1 + \frac{2K_D}{([P]_T - [L]_T - K_D) + \sqrt{([P]_T - [L]_T - K_D)^2 + 4K_D[P]_T}}\right]^{-1} \tag{7.2}$$

The solution of the above equation for K_D yields the following result:

$$K_D = \frac{(f_B - 1)(f_B[L]_T - [P]_T)}{f_B} \tag{7.3}$$

which was used to computed per-peak K_D values for each remodeled spectrum $\mathbf{r}_n$. Finally, the per-peak K_D values were used to compute sample mean and standard deviation K_D values for each ligand. Hit detection was accomplished by comparing per-ligand mean and standard deviation K_D values against a threshold via a Student's t-test, where a resulting p value less than a predefined significant p value was reported as binding.

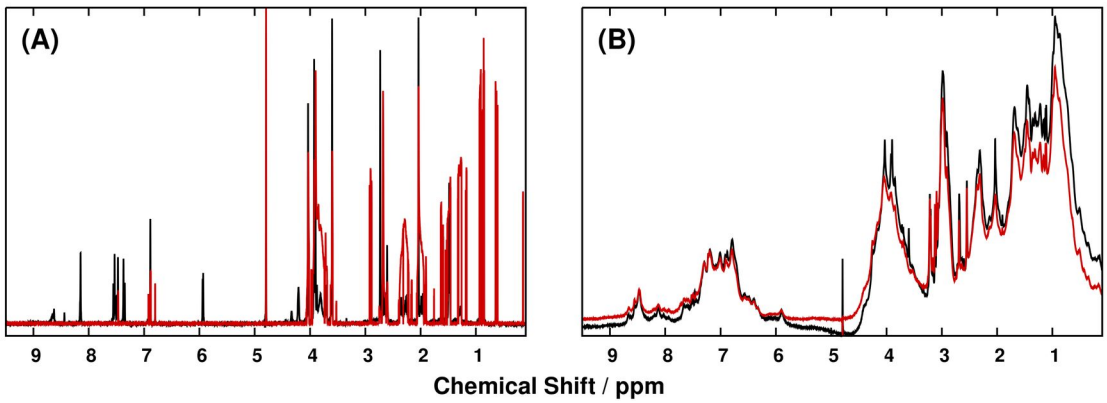

Figure 7.3: Failed Baseline Removal due to Phase Errors.
Example of a failed USSR result, highlighting the impact of phase error during computation and subtraction of the statistical baseline from a screen spectrum. Remodeled peaks (**A**, red) upfield of 4.0 ppm are in fact not true signals, but were generated due to a phase-induced discrepancy between the statistical baseline (**B**, red) and the screen spectrum (**B**, black).

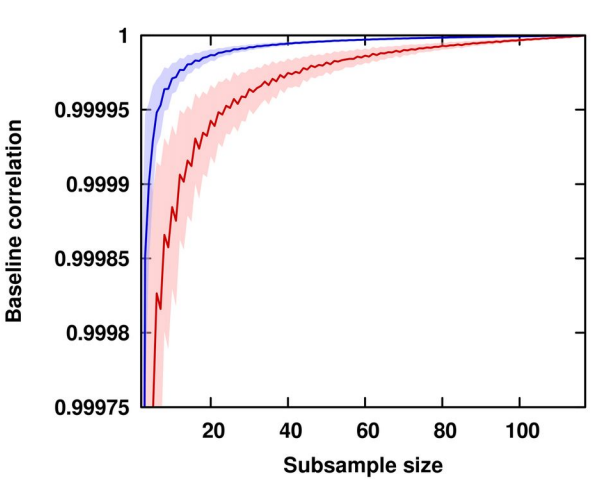

Figure 7.4: Impact of Dataset Size on USSR Statistical Baselines. Correlation between statistical baselines from bootstrap-subsampled datasets of varying size and the original statistical baseline computed from the complete BSA dataset. Lines indicate median correlations, and shaded regions indicate confidence regions of plus or minus one standard deviation, estimated using median absolute deviation. Blue lines and shaded regions indicate values from subsampling the PSC normalized dataset, and red lines and shaded regions indicate values from subsampling the uncorrected dataset.

7.2.5 Analysis of Dataset Size

A small simulation study was conducted to assess the quality of USSR statistical baseline estimates over a range of sample sizes (number of spectral pairs). For sizes from 2 to 116, the BSA dataset was randomly subsampled, without replacement, to produce a smaller dataset. For each resultant dataset, the statistical baseline was estimated, and its Pearson correlation to the true statistical baseline was computed and stored. Over all numbers of spectral pairs in the simulation, the median baseline correlations were computed, and are reported in Figure 7.4.

7.3 Results

From the USSR analysis of ligand binding to BSA, 43 compounds were classified as hits from the library of 456 compounds. All classified hits were determined to bind BSA with at least 1.0 mM affinity ($K_D \leq 0.001$ μM) at a statistical confidence level of 99%. A summary of the hits, along with their estimated K_D and p values, is provided in Table 7.3. Comparison of results from both PSC-corrected and uncorrected USSR datasets reveals that the use of PSC normalization prior to USSR modeling greatly reduces the effective positive rate of statistical hit determination: 195 hits were identified from the PSC-uncorrected spectra. Closer examination of hits identified without PSC correction indicates that USSR failed to fully subtract the statistical baseline from the screen spectra (e.g. Figure 7.3), resulting in residual baseline intensity passing into equation 7.2 during K_D calculation and hit determination. In short, the use of PSC normalization prior to USSR enables more effective baseline subtraction by decreasing both dilution- and phase-related protein baseline intensity variation in collected ^{1}H NMR spectra (Figure 7.5). Baseline estimates obtained by collecting a spectrum of pure protein will suffer from the same phase-induced variation, which

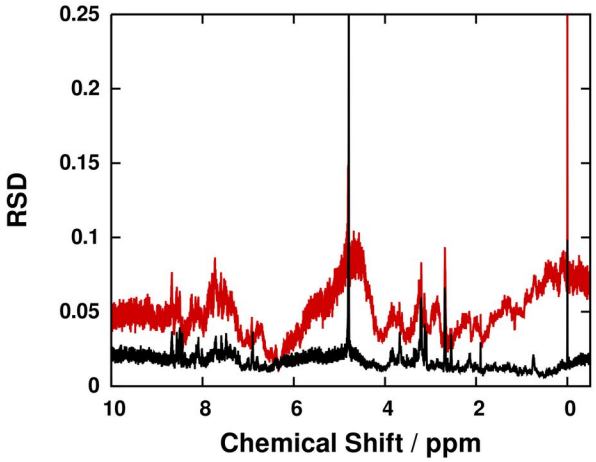

Figure 7.5: Impact of PSC on USSR Statistical Baselines.
Relative standard deviations (RSDs) of the statistical baselines computed before (red) and after (black) phase-scatter correction, which substantially decreases inter-sample variability of the protein baseline signals.

would also increase the false positive rate during hit determination. The introduced combination of PSC and USSR provides a more reliable means of baseline identification, without the need for collection of a free protein reference spectrum.

Cursory analysis of the robustness of the USSR statistical baseline during random subsampling of the BSA dataset indicated that the PSC/USSR methodology can reliably operate at very low dataset sizes (i.e. 10–20 spectral pairs). Pearson correlations between true and subsampled baselines did not appreciably decrease even after harsh subsampling (Figure 7.3), and correlations computed from PSC-normalized data maintained significantly higher values than those from non-normalized data. While it would be possible to obtain a statistical baseline from fewer than ten spectral pairs, this is not recommended, as it will decrease the effectiveness of the Bonferroni-corrected t-test that USSR performs during remodeling. Therefore, as a general rule of thumb, PSC/USSR analyses may be performed on high-throughput screening datasets having as few as ten spectral pairs, and higher sample sizes only serve to further increase the reliability of remodeled results.

7.4 Discussion and Conclusions

While the saturation transfer difference (STD) NMR experiment [11] is a popular choice for ligand-based NMR affinity screens, a 1D ^{1}H NMR spectrum requires only a few seconds to acquire, making it an ideal choice for high-throughput screening. STD experiments require significantly longer acquisition times (upwards of hours) in order to acquire difference spectra with sufficient signal-to-noise to reduce false negatives. A particular strength of STD is the minimal amount of protein needed per experiment, making it practical to screen a reasonably large chemical library (upwards of thousands

of compounds) with only a few milligrams of protein. Through a judicious choice of protein and ligand concentrations coupled with the use of cryoprobes and high magnetic fields, it is also possible to minimize protein requirements in 1D ^{1}H line-broadening screens. While STD experiments still tend to require less protein than line-broadening experiments, the higher false positive rate of STD screening easily negates any advantages of minimal protein usage. This high false positive rate arises due to the tendency of STD experiments to emphasize weak binding affinities commonly encountered during aggregation and nonspecific binding [7, 9].

NMR line-broadening experiments take advantage of the molecular-weight dependence of T_2 relaxation and the resultant measurable difference in line-widths between proteins and the compounds in a screening library [6]. Upon binding a protein target, the ^{1}H NMR resonances of a compound will broaden significantly or even disappear. In principal, this spectral broadening is easily observable and binding is readily identified. In practice, background signals from the protein can confound the data analysis. This background interference increases with the size and concentration of the protein and leads to an increase in false negative rates. Apparent line-width differences between free and bound ligands *also* increase with protein size and concentration, making the optimal experimental conditions for NMR line-broadening screens exactly the same conditions which confound manual interpretation. Clearly, the ability to accurately remove the protein background from an NMR line-broadening experiment will improve both the utility and reliability of the technique, especially at relatively high protein-ligand concentration ratios where binding is more apparent.

By removing interfering protein baseline signals, USSR provides a straightforward means of visually or computationally analyzing screening results. In fact, the outcome of applying the USSR method to an extremely challenging and atypical test case is rather dramatic: an NMR line-broadening screen of BSA against a chemical library of 456 compounds identified 43 binders, despite the BSA background signals completely obscuring the ligand spectral features. An example screening result of tolazamide, dimethyl 4-methoxyisophthalate, 1,7-dimethylxanthine and oxolinic acid against BSA is illustrated in Figure 7.2. Removal of the interfering protein statistical baseline from the screen spectrum (Figure 7.2B) yielded a high-quality pseudo-spectrum of the ligand mixture in the presence of BSA. Overlaying the remodeled NMR spectrum with the free ligand mixture spectrum indicated that the two spectra were essentially identical for the non-binding ligands (Figure 7.2A). Only dimethyl 4-methoxyisophthalate, which binds BSA, exhibited any difference after remodeling. The USSR method of baseline estimation and subtraction is expected to perform equally well under

any conditions where a common, highly reproducible spectral feature exists within a dataset. The presented application of PSC/USSR to high-throughput protein-ligand affinity screening is but one example of its potential uses.

However, reliable identification of the protein baseline from screening data requires highly reproducible sample preparation, data collection and processing. The last of these requirements is met by the use of phase-scatter correction prior to remodeling, which brings protein baselines from all spectra into closer agreement with each other and minimizes the number of false hits identified during analysis. It is important to note that PSC is only an effective pre-treatment for USSR when protein baseline signals are of comparable intensity to ligand signals. PSC normalization is designed to maximize statistical agreements *between* spectra by phase and normalization correction, and its use of the ℓ_2 norm as a criterion for 'agreement' implies that higher-intensity features will be preferentially corrected. Thus, PSC achieves the best results prior to USSR when protein signals are a major spectral component, as is the case when protein-ligand concentration ratios are near or greater than unity. In effect, the combined use of PSC and USSR expands the range of protein-ligand concentration ratios which may be probed by ^{1}H line-broadening experiments for the purposes of high-throughput screening.

Finally, it cannot be under-emphasized that the single-point K_D computations employed by USSR during statistical hit determination are only order-of-magnitude estimates of the true dissociation rates, and can carry significant systematic and random errors. In particular, the fraction of bound ligand – and by extension, the dissociation constant – depends exquisitely on the estimated free and bound ligand line widths, v_F and v_B. Thus, any imprecision in the line width estimates will propagate into a systematic bias in the final dissociation constants. If required, verification of initial hits may be achieved to higher accuracy via multiple-point estimation of the K_D through linear or nonlinear least squares [21].

An implementation of the USSR algorithm is available in open-source GNU Octave code as a part of the MVAPACK toolbox [23].

Table 7.1: Results of the USSR Analysis of Ligand Binding to BSA.

Compound	K_D **(nM)**	p
Thiamine hydrochloride	0.143	$\leq 10^{-6}$
5-azacytidine	0.072	$\leq 10^{-6}$
Nadolol	0.142	$\leq 10^{-6}$
N-succinyl-Ala-Ala-Pro-Phe p-nitroanilide	0.090	$\leq 10^{-6}$
Timolol maleate	0.092	$\leq 10^{-6}$
5-phenylvaleric acid	0.000	$\leq 10^{-6}$
Astemizole	0.000	$\leq 10^{-6}$
Leflunomide	0.041	$\leq 10^{-6}$
Gliotoxin	0.061	$\leq 10^{-6}$
2-aminofluorene	0.141	0.000001
Mordant orange 1	0.162	0.000020
Indomethacin	0.239	0.000027
Diminazene aceturate	0.186	0.000034
Flavanone	0.188	0.000065
(-)-arctigenin	0.200	0.000101
Meclofenamic acid sodium salt	0.154	0.000188
2-acetamidophenol	0.198	0.000202
Bromocresol green	0.198	0.000266
Mycophenolic acid	0.203	0.000341
7-deazaguanine	0.213	0.000466
Camptothecin	0.183	0.000468
Flavone	0.208	0.000582
Naproxen	0.211	0.000653
2,6-diisopropylphenol	0.220	0.000816
Z-L-phenylalanine	0.230	0.001076
Cromolyn sodium salt	0.226	0.001080
Methotrexate, (+)-ametthopterin	0.237	0.001106
Cinoxacin	0.231	0.001224
Dimethyl 4-methoxyisophthalate	0.218	0.001315
8-methoxypsoralen	0.237	0.002242
L-ornithine hydrochloride	0.337	0.002326
Aminophylline hydrate	0.246	0.002633
Captopril	0.277	0.002781
Ebselen	0.246	0.002854
2-aminophenol	0.394	0.003476
Myristic acid	0.271	0.004925
Sulindac sulfide	0.269	0.005659
Phenylpyruvic acid	0.282	0.006867
Diclofenac sodium salt	0.284	0.007310
Prednisolone	0.286	0.007693
Alaproclate hydrochloride	0.281	0.007706
Phenylmethanesulfonyl fluoride	0.265	0.008433
Methyl 4-hydroxyphenylacetate	0.286	0.008845

7.5 References

[1] C. Dalvit, P. Pevarello, M. Tato, M. Veronesi, A. Vulpetti, and M. Sundstrom. Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *Journal of Biomolecular Nmr*, 18(1):65–68, 2000.

[2] O. J. Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52, 1961.

[3] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[4] A. E. Ferentz and G. Wagner. NMR spectroscopy: a multifaceted approach to macromolecular structure. *Quarterly Reviews of Biophysics*, 33(1):29–65, 2000.

[5] P. J. Hajduk and J. Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery*, 6:211–219, 2007.

[6] P. J. Hajduk, E. T. Olejniczak, and S. W. Fesik. One-dimensional relaxation- and diffusion-edited NMR methods for screening compounds that bind to macromolecules. *Journal of the American Chemical Society*, 119(50):12257–12261, 1997.

[7] M. J. Harner, A. O. Frank, and S. W. Fesik. Fragment-based drug discovery using NMR spectroscopy. *Journal of Biomolecular NMR*, 56(2):65–75, 2013.

[8] T. L. Hwang and A. J. Shaka. Water Suppression That Works – Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic Resonance*, 112(2):275–279, 1995.

[9] C. A. Lepre. Practical aspects of NMR-based fragment screening. *Methods Enzymol*, 493:219–239, 2011.

[10] C. a. Lepre, J. M. Moore, and J. W. Peng. Theory and applications of NMR-based screening in pharmaceutical research. *Chemical Reviews*, 104(8):3641–3675, 2004.

[11] M. Mayer and B. Meyer. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *Journal of the American Chemical Society*, 123(25):6108–6117, 2001.

[12] K. A. Mercier, M. Baran, V. Ramanathan, P. Revesz, R. Xiao, G. T. Montelione, and R. Powers. FAST-NMR: Functional annotation screening technology using NMR spectroscopy. *Journal of the American Chemical Society*, 128:15292–15299, 2006.

[13] K. A. Mercier, M. D. Shortridge, and R. Powers. A Multi-Step NMR Screen for the Identification and Evaluation of Chemical Leads for Drug Discovery. *Combinatorial Chemistry and High Throughput Screening*, 12(3):285–295, 2009.

[14] A. Muller, R. M. MacCallum, and M. J. E. Sternberg. Structural characterization of the human proteome. *Genome Research*, 12(11):1625–1641, 2002.

[15] B. D. Nguyen, X. Meng, K. J. Donovan, and A. J. Shaka. SOGGY: Solvent-optimized double gradient spectroscopy for water suppression. A comparison with some existing techniques. *Journal of Magnetic Resonance*, 184(2):263–274, 2007.

[16] M. Pellecchia, D. S. Sem, and K. Wuthrich. NMR in drug discovery. *Nature Reviews Drug Discovery*, 1:211–219, 2002.

[17] R. Powers. Advances in nuclear magnetic resonance for drug discovery. *Expert Opinions in Drug Discovery*, 4(10):1077–1098, 2009.

[18] R. Powers, J. C. Copeland, K. Germer, K. A. Mercier, V. Ramanathan, and P. Revesz. Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design. *Proteins: Struct Funct Bioinformatics*, 65(1):124–135, 2006.

[19] R. Powers, J. C. Copeland, J. L. Stark, A. Caprez, A. Guru, and D. Swanson. Searching the protein structure database for ligand-binding site similarities using CPASS v.2. *BMC Research Notes*, 4(17):1–15, 2011.

[20] R. Powers, K. a. Mercier, and J. C. Copeland. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discovery Today*, 13(3-4):172–179, 2008.

[21] M. D. Shortridge, D. S. Hage, G. S. Harbison, and R. Powers. Estimating protein-ligand binding affinity using high-throughput screening by NMR. *Journal of Combinatorial Chemistry*, 10(6):948–958, 2008.

[22] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274(5292):1531–1534, 1996.

[23] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

Chapter 8
Generalized Adaptive Intelligent Binning of Multiway Data

The art of doing mathematics consists in finding that special case which contains all the germs of generality.

– David Hilbert

8.1 Introduction

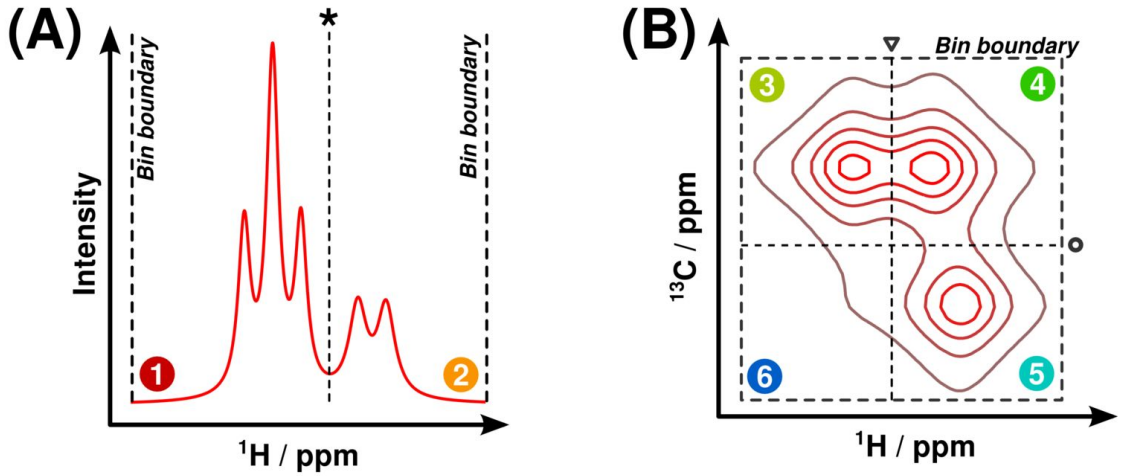

Figure 8.1: Generalization of Adaptive Intelligent Binning.
(**A**) In the one-dimensional case, the bin containing regions 1 and 2 is optimally subdivided (*asterisk*) when the sum of the objective values in regions 1 and 2 is maximal and greater than the original bin's objective value. (**B**) In the D-dimensional case, there are now D possible dimensions along which an optimal subdivision may exist. The optimal subdivision along the ^{1}H dimension (*triangle*) occurs when the sum of the objective values in regions 3+6 and 4+5 is maximal and greater than that of the original bin. Similarly, the optimal subdivision along the ^{13}C dimension (*circle*) occurs when the sum of the objective values in regions 3+4 and 5+6 is maximal and above the original bin objective. A comparison between all possible optimal subdivisions along all dimensions yields the best possible subdivision (^{13}C, *circle*).

By and large, the phase "NMR metabolic fingerprinting" implies the use of one-dimensional (1D) ^{1}H NMR spectroscopic methods, due in no small part to the ease and speed of 1D data collection and

the large natural abundance of NMR-active protons found in metabolomics samples [20, 33]. Before processed spectra are submitted to PCA or PLS for modeling, they are often subdivided into bins to simplify multivariate analyses. Spectral binning, introduced and described in detail in Chapter 3, reduces the dimensionality of a data matrix and masks chemical shift variability between samples at the expense of decreased model interpretability: any given bin in a 1D ^{1}H NMR spectral dataset may contain several overlapped signals from multiple distinct metabolites [1]. Thus, without utilizing computationally intensive methods of deconvolution to tease apart signal contributions from individual metabolites [4, 36], the resulting fingerprint from a binned 1D dataset is usually limited to high-level inference about metabolic trends.

By leveraging the connectivities between ^{1}H and ^{13}C nuclei in metabolites, two-dimensional (2D) heteronuclear NMR methods reduce spectral overlap by spreading ^{1}H information over a second (^{13}C) chemical shift dimension [23]. Heteronuclear single quantum coherence (HSQC) experiments are commonly performed in NMR metabolic profiling studies, and provide an NMR singlet or multiplet for each directly bonded ^{1}H–^{13}C pair in the sample. Developments in NMR hardware and acquisition techniques have brought natural abundance ^{1}H–^{13}C HSQC experiment times down to values compatible with high-throughput metabolic fingerprinting studies [24, 26]. However, multivariate analysis of 2D NMR datasets is still a nontrivial undertaking that requires either vectorization [16], which breaks the inherent structure of the data, or the use of multilinear factorizations [21, 22], which are more computationally intensive and difficult to cross-validate.

Spectral binning is another potential means of preparing 2D NMR datasets for multivariate analysis that holds several advantages over binning 1D spectra. First, multiple integration of bins maps each spectrum to an observation vector regardless of its original dimensionality, allowing bilinear PCA and PLS algorithms to be used without concern for loss of the inherent structure of the data. Second, binning of 2D spectral data yields more well-conditioned data matrices than simple vectorization. Finally, because signals are better resolved in 2D spectra, each bin contains substantially fewer signals from distinct metabolites. Multiple different algorithms have been developed to bin 1D NMR data [3, 2, 7, 9, 27], and the use of uniform binning on 2D NMR data has also been reported [29]. However, at the time of this writing, no methods exist to *intelligently* bin multidimensional data for use in multivariate analyses. This motivated the development of a generalization of Adaptive Intelligent (AI) binning [9] to spectral data of any dimensionality, called Generalized Adaptive Intelligent (GAI) binning (Figure 8.1).

8.2 Theory

8.2.1 AI-binning

Generalized AI-binning (GAI-binning) is a logical extension of AI-binning to two or more dimensions. In the AI algorithm (Figure 8.1A), bins are recursively subdivided until a stopping criterion or minimum bin width is reached [9]. For a 1D dataset containing N spectra, the following objective function is used to assess the quality of each bin:

$$V_b = \frac{1}{N}\sum_{n=1}^{N}\left[(max_{n,b} - I_{n,b,1})(max_{n,b} - I_{n,b,end})\right]^{\frac{R}{2}} \tag{8.1}$$

where $max_{n,b}$ is the maximum intensity inside the bin b in spectrum n, and $I_{n,b,1}$ and $I_{n,b,end}$ are the bin edge intensities. The exponent R in the AI objective function is referred to as a "resolution parameter", which offers a means of tuning the binning result based on signal-to-noise and peak resolution of a dataset. The replacement of R with $\frac{R}{2}$ in the exponent of equation 6.1, enables a slightly modified interpretation of each summed term in the AI objective function as a relaxed form of a geometric mean of the differences between the bin edge intensities and the maximum bin intensity. At each subdivision step, new bin edges are chosen to maximize the combined (summed) objective values of the two resulting bins over the objective value of the original bin. If no bin subdivision exists with a combined objective value greater than that of the original bin, recursive subdivision within that bin is terminated, and the AI algorithm terminates once all bins may no longer be subdivided.

8.2.2 GAI-binning

In two or more dimensions, the set of bin boundary points expands to include all points that lie on the edges (or faces, hyperfaces, etc.) of the bin. By denoting the set of all edge points in bin b as E_b, a new objective function may be constructed:

$$V_b = \frac{1}{N}\sum_{n=1}^{N}\left[\prod_{e\in E_b}(max_{n,b} - I_e)\right]^{\frac{R}{||E_b||}} \tag{8.2}$$

Thus, the GAI algorithm computes the "relaxed" geometric mean of the differences between the bin maximum and all points on the boundary. In the case of one-dimensional data, it is apparent that equation 6.2 reduces to equation 6.1, and GAI-binning operates identically to AI-binning. As dimensionality increases, the risk of floating-point overflow or underflow increases due to the larger

bin edge set E_b. To avoid this, the following "log-objective" may be used in lieu of equation 6.2:

$$V_{b,ln} = \frac{R}{N||E_b||} \sum_{n=1}^{N} \sum_{e \in E_b} \ln(max_{n,b} - I_e) \tag{8.3}$$

Like AI-binning, GAI-binning initializes a bin around the entire dataset and proceeds to recursively subdivide each bin until a minimum bin size is reached or no bin may be divided to yield an increase in the objective value. Because the number of ways to subdivide each bin increases with dimensionality, all possible dimensions are tested, and the new bin boundary that maximizes the objective over all possible subdivision dimensions is selected (Figure 8.1B). Therefore, the GAI algorithm may be considered a form of binary space partitioning (BSP) which limits its partition hyperplanes to lying orthogonally to the basis vectors of the coordinate system [8].

8.2.3 Noise Bin Elimination

It is important that noise bins be removed from the data matrix prior to multivariate analysis, as their presence is known to negatively impact the interpretability and reliability of multivariate models [15, 5]. Because the integration of a noisy space of increasing dimensionality (i.e. double or triple integration) results in a random variable having a similarly increasing variance, the importance of noise removal is compounded in multidimensional binning. Therefore, a noise bin removal step based on spectral intensity was added to the GAI algorithm. A running mean and variance calculation was performed to estimate the noise floor of each spectrum. The initial mean μ_n and standard deviation σ_n of the noise were computed using the first 32 points on one edge of the spectrum, which were assumed to contain only baseline noise. Every other data point was then classified as signal or noise based on whether its intensity exceeded the current running noise floor, $\mu_n + 3\sigma_n$. Upon inclusion of a new noise data point, the mean and standard deviation of the noise were appropriately updated. Once the estimated noise floor was determined for each spectrum in the dataset, a threshold for bin removal was computed as the median noise floor of all the spectra:

$$I_{th} = \text{med}_n(\mu_n + k\sigma_n) \tag{8.4}$$

where k is a user-selectable parameter to adjust the noise threshold. Only bins whose maximum intensity fell above I_{th} were retained in the final data matrix.

8.3 Materials and Methods

8.3.1 Human Liver Dataset

Two independently collected ^{1}H–^{13}C HSQC NMR datasets from ongoing metabolomics studies were used as test cases for the GAI-binning algorithm. For the first dataset, twenty-four 1.0 mL samples of SK-Hep1 human liver cells were provided for metabolic fingerprinting, half of which were treated with 50 μM tetrathiomolybdate (TTM). The cells were extracted into 80:20 methanol:water to collect the water-soluble metabolites, spun in a rotary evaporator for two hours, lyophilized at $-50°$C and 0.02 mBar for 24 hours, and finally redissolved in 600 μL of 50.0 mM phosphate buffer in 99.8% D_2O (Isotec, St. Louis, MO) adjusted to pH 7.4. The redissolved, pH-adjusted samples were then collected into NMR tubes.

Experiments were collected on a Bruker Avance III HD 700 MHz spectrometer equipped with a 5 mm inverse quadruple-resonance (^{1}H, ^{13}C, ^{15}N, ^{31}P) cryoprobe with cooled ^{1}H and ^{13}C channels and a z-axis gradient. A Bruker SampleJet and ICON-NMR were used to automate NMR data collection. A 2D gradient-enhanced ^{1}H–^{13}C HSQC with improved sensitivity [25, 18] (*hsqcetgpsi*) was collected for each sample. Spectra were collected with 4 scans and 16 dummy scans over a uniform Nyquist grid of 512 and 64 complex points along the ^{1}H and ^{13}C dimensions, respectively. Spectral windows were set to 3,285 $\pm$ 4,545 Hz along ^{1}H and 12,677 $\pm$ 14,620 Hz along ^{13}C. All spectra were collected at a sample temperature of 298.0 K.

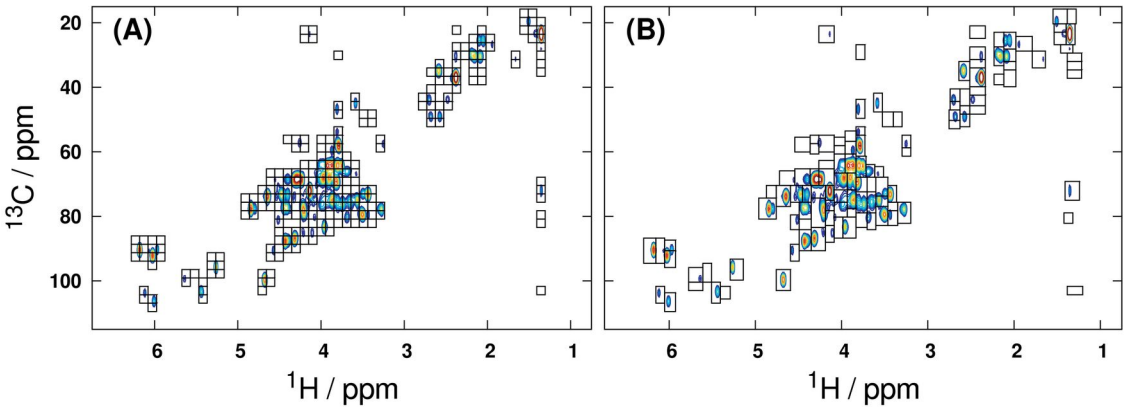

Figure 8.2: Binned Liver Dataset.
Processed ^{1}H–^{13}C HSQC mean spectrum of the liver data tensor, with overlaid uniform (**A**) and GAI (**B**) bin boundaries.

8.3.2 Mouse Embryonic Fibroblast Dataset

A second set of samples from kinase suppressor of Ras 1 (KSR1) knockout mouse embryonic fibroblast (MEF) cells was also provided to generate a test ^{1}H–^{13}C HSQC dataset for GAI-binning. For this second dataset, ten cell samples from $ksr^{-/-}$ MEFs and ten samples from KSR1-rescued $ksr^{-/-}$ MEFs were used to produce metabolite extracts. The cells were washed, extracted into 80:20 methanol:water, spun in a rotary evaporator, lyophilized and redissolved according to the procedures used to extract metabolites from the liver cell samples.

Experiments were collected on a Bruker Avance DRX 500 MHz spectrometer equipped with a 5 mm inverse triple-resonance (^{1}H, ^{13}C, ^{15}N) cryoprobe with a z-axis gradient. A Bruker BACS-120 sample changer and ICON-NMR software were used to automate data collection. A 2D gradient-enhanced ^{1}H–^{13}C HSQC (*hsqcetgp*) was collected for each sample. Spectra were collected with 128 scans and 16 dummy scans over a uniform grid of 1024 and 32 complex points along the ^{1}H and ^{13}C dimensions, respectively. Spectral windows were set to 2,359 ± 2,367 Hz along ^{1}H and 8,174 ± 8,803 Hz along ^{13}C. All spectra were collected at a sample temperature of 293 K.

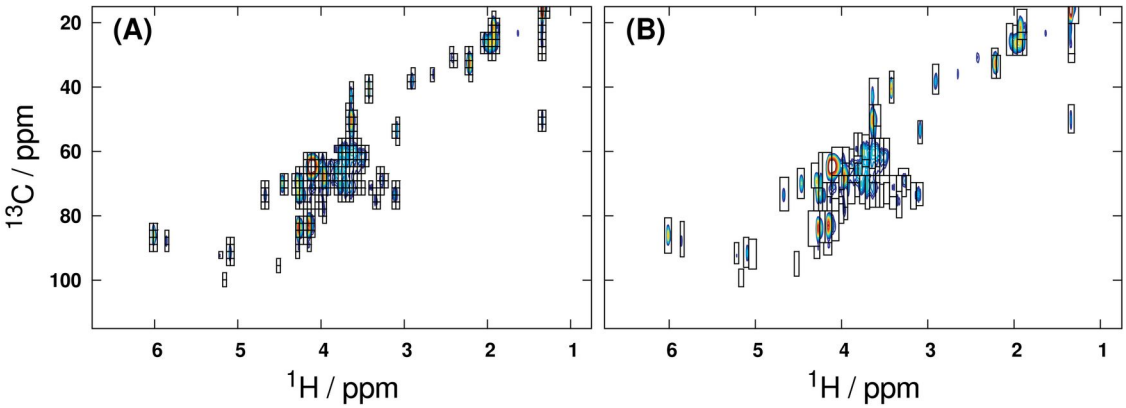

Figure 8.3: Binned Fibroblast Dataset.
Processed ^{1}H–^{13}C HSQC mean spectrum of the MEF data tensor, with overlaid uniform (**A**) and GAI (**B**) bin boundaries.

8.3.3 NMR Processing and Multivariate Analysis

All processing, treatment and statistical modeling were performed in GNU Octave 3.6 [12] using routines currently available in the MVAPACK toolbox for NMR chemometrics [34], discussed in Chapter 4. The 2D raw serial files were loaded [10], apodized with a squared-sine window, zero-filled once along ^{1}H and twice along ^{13}C, and Fourier-transformed. Spectra from the liver cell extracts

were manually phase-corrected and cropped (1.0 – 6.6 ppm along ^{1}H; 16 – 112 ppm along ^{13}C), and spectra from the MEF extracts were similarly phase corrected and cropped (1.25 – 6.2 ppm along ^{1}H; 8 – 102 ppm along ^{13}C). Both uniform and GAI-binning were performed on each data tensor using minimum ^{1}H and ^{13}C bin widths of 0.025 and 2.5 ppm, respectively, and a GAI resolution parameter of 0.1. Binned regions identified to be less intense than three times the standard deviation of the spectral noise ($k = 3$) were removed after binning. The mean spectra of the entire processed liver and MEF datasets, superimposed with bins identified by both uniform and GAI-binning, are shown in Figure 8.2 and Figure 8.3.

The applicability of GAI-binning to bilinear factorizations was demonstrated by modeling the data tensors using both PCA and OPLS-DA. For PCA modeling of the data, the spectral regions identified by each binning method were doubly integrated. Scores and loadings were then calculated using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [17]. Internal leave-one-out cross-validation (LOOCV) of each computed PCA model was performed to yield model fit (R^2_X) and predictive ability (Q^2) statistics [19, 14]. For OPLS-DA, spectral data points within the identified bins were vectorized row-wise into a data matrix as previously described [16]. During vectorization, all data points within each binned region are stacked into an observation vector, and data points not within bins are excluded. The use of vectorization prior to supervised modeling facilitates the creation of backscaled pseudospectral OPLS loadings, which hold greater ease of interpretation over binned loadings [32]. Modeling by an OSC-filtered NIPALS algorithm [28] and 100 rounds of seven-fold Monte Carlo cross-validation (MCCV) [35] were performed to compute data fit (R^2_X), response fit (R^2_Y) and model predictive ability (Q^2) statistics. The binned data matrices produced via double integration were also subjected to OPLS-DA modeling in the same manner as the vectorized data. All OPLS-DA models were further validated using CV-ANOVA [13] and 1,000 iterations of response permutation testing [31] to rigorously ensure model reliability. Backscaled predictive OPLS loadings were computed from the vectorized bins according to previously published works [6, 16]. During backscaling, OPLS loading vectors were scaled by the inverse of their original Pareto scaling coefficients and then unstacked into a two-dimensional pseudospectrum using bin information. Data points not included in the vectorized loadings were set to zero in the backscaled pseudospectrum. All data matrices were normalized using Probabilistic Quotients (PQ) [11] and then Pareto scaled [30] prior to modeling.

8.4 Results and Discussion

Processing of the liver extract spectra yielded a real data tensor of 24 ^{1}H–^{13}C HSQC spectra having 442×149 points each, and processing of the fibroblast spectra yielded a tensor of 17 spectra having 1,071×172 real data points each. The observation counts (N), variable counts (K) and PCA/OPLS cross-validation statistics (R^2, Q^2) for each dataset and variable reduction method are summarized in Table 8.1. Further validation results from the OPLS models, all of which indicate varying degrees of high model reliability, are also summarized in Table 8.2. Through examination of the variable counts within Table 8.1, it is readily apparent that GAI-binning is dramatically more effective than uniform binning at discriminating between signal and noise regions within spectral data. On average, GAI-binning segmented each data tensor into less than half the number of bins produced by uniform binning, and produced PCA models with markedly higher R^2_X and Q^2 statistics. Moreover, even with the greatly reduced variable counts produced by GAI-binning relative to uniform binning, the OPLS Q^2 statistics between the two methods are statistically indistinguishable. In fact, the variable counts resulting from GAI-binning these third-order tensors are substantially lower than the few hundred variables typically produced by binning *one*-dimensional spectra. Resulting scores from PCA modeling of the GAI-binned liver data tensor are shown in Figure 8.4.

Table 8.1: Data Matrices and PCA/OPLS Model Statistics.

		Integration					**Vectorization**		
		PCA			OPLS		OPLS		
		K	R^2_X	Q^2	R^2_Y	Q^2	K	R^2_Y	Q^2
Liver	Unif.	248	0.82	0.71	0.993	0.938 ± 0.002	11,160	0.993	0.929 ± 0.003
$N = 24$	GAI	113	0.89	0.75	0.991	0.928 ± 0.003	10,474	0.994	0.933 ± 0.003
MEF	Unif.	334	0.48	0.40	0.994	0.974 ± 0.004	18,348	0.994	0.963 ± 0.005
$N = 17$	GAI	93	0.71	0.56	0.994	0.973 ± 0.005	18,789	0.996	0.962 ± 0.006

Table 8.2: OPLS-DA Cross Validation p-values.

		Integration		**Vectorization**	
		Permutation	CV-ANOVA	Permutation	CV-ANOVA
Liver	Unif.	< 0.001	3.24×10^{-11}	< 0.001	4.70×10^{-11}
$N = 24$	GAI	< 0.001	3.34×10^{-10}	< 0.001	9.74×10^{-11}
MEF	Unif.	< 0.001	3.56×10^{-10}	< 0.001	1.73×10^{-9}
$N = 24$	GAI	< 0.001	1.37×10^{-9}	< 0.001	2.34×10^{-9}

Backscaled predictive OPLS-DA loadings of the vectorized ^{1}H–^{13}C HSQC spectral data tensors

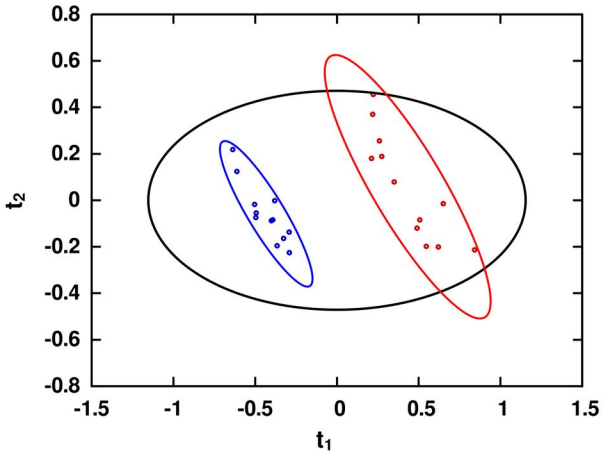

Figure 8.4: PCA Scores of a GAI-binned Tensor.
Principal component analysis scores resulting from modeling the GAI-binned ^{1}H–^{13}C HSQC liver data matrix, indicating a high degree of separation between experimental groups. Model R^2_X and Q^2 were 0.68 and 0.64 for the first principal component (t_1) and 0.12 and 0.09 for the second (t_2). Class separations of this magnitude are readily achievable using data matrices generated by GAI-binning, due in large part to the low variable counts it generally produces.

(Figure 8.5) lend further support for the use of multidimensional binning in metabolic fingerprinting studies. Even when vectorization is performed in place of integration to produce a data matrix, binning offers an effective means of variable selection: only 10,474 of 65,858 variables (16%) were retained when GAI-binning was used as a pre-filter prior to modeling the liver data. A similar reduction was observed in the fibroblast dataset, where GAI-binning retained 18,789 of 184,212 total variables for a 90% reduction in dimensionality. These substantially reduced variable counts offered by binning translate to more well-conditioned bilinear modeling problems. As the dimensionality of the input dataset is increased further, the reductions in variable count afforded by multidimensional binning are expected to become even more dramatic. While the variable counts produced by vectorization of uniformly binned data tensors are comparable to those from GAI-binning, it is critical to recognize that the uniformly binned regions contain more noise data points than their GAI-binned counterparts, and thus offer a less efficient dimensionality reduction.

Spectral regions produced by GAI-binning (Figure 8.2) demonstrate several important properties of the combined binning and noise removal processes. Because t_1 noise and truncation artifacts yield phase-incoherent negative spectral excursions after Fourier transformation, "unrelaxed" GAI-binning ($R = 1$) tends to preferentially subdivide near such regions, producing elongated bins along the F_1 dimension. Decreasing the resolution parameter from its maximum value shrinks these bins to contain only true signals. Thus, an objective rule for determining an optimal resolution parameter during binning is to decrease R until all bins shrink to contain a minimal amount of noise. Once an optimal resolution parameter has been identified, a suitable noise threshold (k) must be determined such that all noise bins are removed without loss of bins containing weak signals. However, once R and k have been determined for a given set of experimental conditions, they may be applied

during GAI-binning to any data collected at later times under the same conditions to achieve ideal results. The selections of resolution parameter ($R = 0.1$) and noise threshold ($k = 3$) made in this work were identified according to the above criteria through a manual visual examination of the binning results, but it is conceivable that objective metrics of these criteria could be constructed that facilitate automated determination of these parameters.

Finally, like AI-binning, the execution time of GAI-binning scales quadratically with the number of spectral data points, and scales approximately linearly with both the number of spectral dimensions and the number of observations. Typical run times for binning two-dimensional datasets range from seconds to a few minutes, depending mostly on the data point count. Thus, while zero-filling may be used to increase the digital resolution of data being input into GAI-binning, it should be applied sparingly to avoid unnecessarily long computation times during bin region determination.

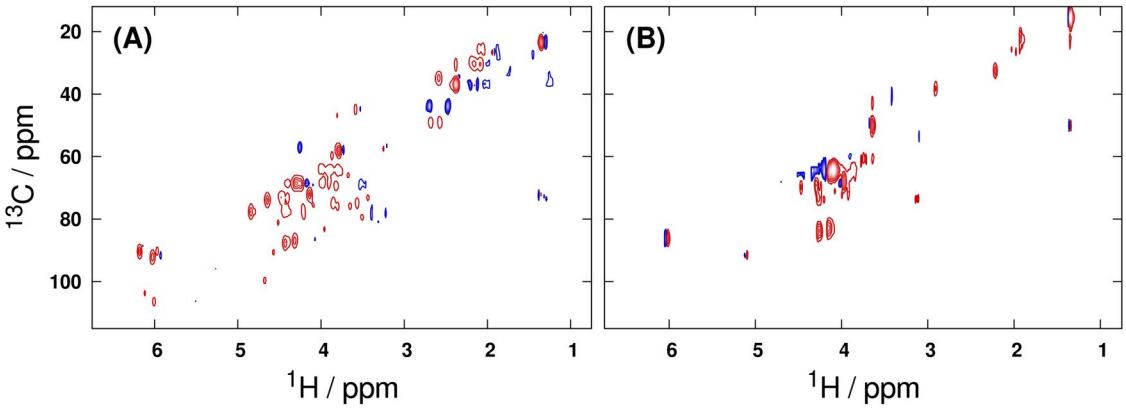

Figure 8.5: Pseudospectral HSQC Loadings.
Backscaled full-resolution pseudospectral loadings from OPLS-DA modeling of the GAI-reduced (**A**) liver and (**B**) fibroblast ^{1}H–^{13}C HSQC data tensors. Positive and negative loadings are represented by red and blue contours, respectively.

8.5 Conclusions

Generalized Adaptive Intelligent binning is a logical extension of the previously described Adaptive Intelligent binning algorithm [9] to multidimensional datasets, and provides a model-free alternative to peak-fitting and peak-picking as a means of variable selection in multivariate analyses. Furthermore, GAI-binning is a more intelligent method to extract signal regions from multidimensional spectral data tensors than uniform binning, and may be used to generate very low-dimensionality data matrices via multiple integration or efficiently noise-filtered data matrices via vectorization.

The C++ implementations of 1D and 2D GAI-binning used in this work are freely available as part of the MVAPACK software package [34] introduced in Chapter 5.

8.6 Permutation Test Results

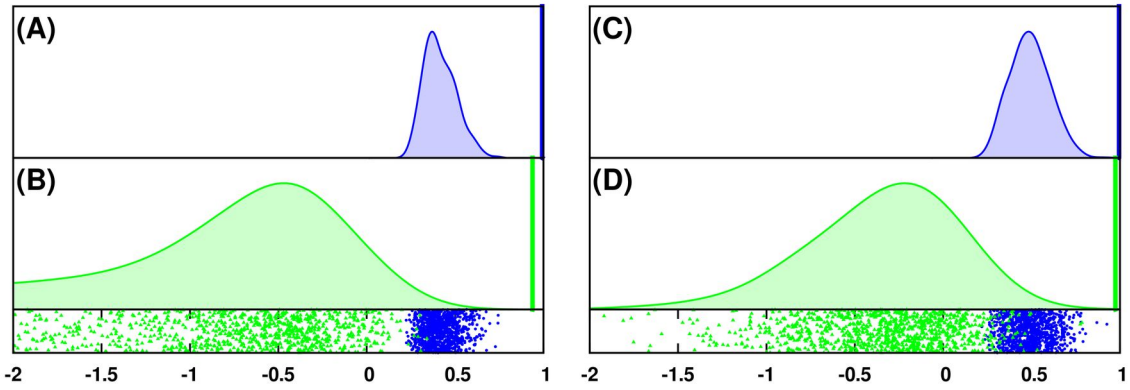

Figure 8.6: Response Permutation Test: Uniform integration.
Response permutation test results for OPLS-DA models from the uniformly binned (integrated) liver (**A**, **B**) and fibroblast (**C**, **D**) data tensors. Model fit (R^2_Y) statistics (**A**, **C**) are shown in blue, and model predictive ability (Q^2) statistics (**B**, **D**) are shown in green. True values of R^2_Y and Q^2 are represented by vertical bars, and null distributions are computed through kernel density estimation of the values from permutation. Scatter plots of the permutation (null) R^2_Y and Q^2 statistics are shown in the lower panes.

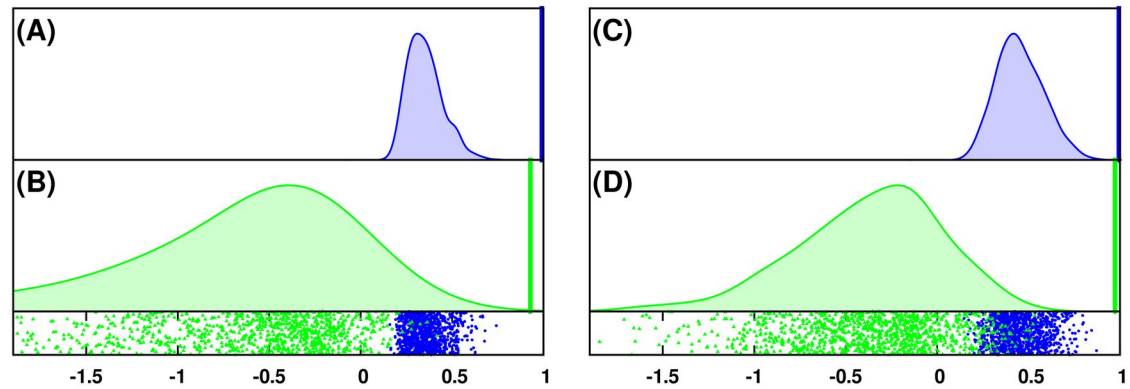

Figure 8.7: Response Permutation Test: GAI-integration.
Response permutation test results for OPLS-DA models from the GAI-binned (integrated) liver (**A**, **B**) and fibroblast (**C**, **D**) data tensors. See the caption of Figure 8.6 for a complete description of the figure contents.

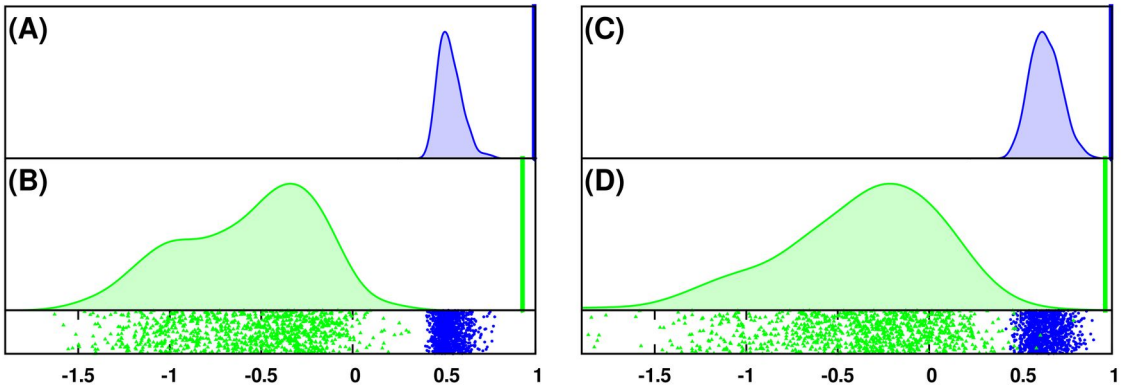

Figure 8.8: Response Permutation Test: Uniform vectorization.
Response permutation test results for OPLS-DA models from the uniformly binned (vectorized) liver (**A**, **B**) and fibroblast (**C**, **D**) data tensors. See the caption of Figure 8.6 for a complete description of the figure contents.

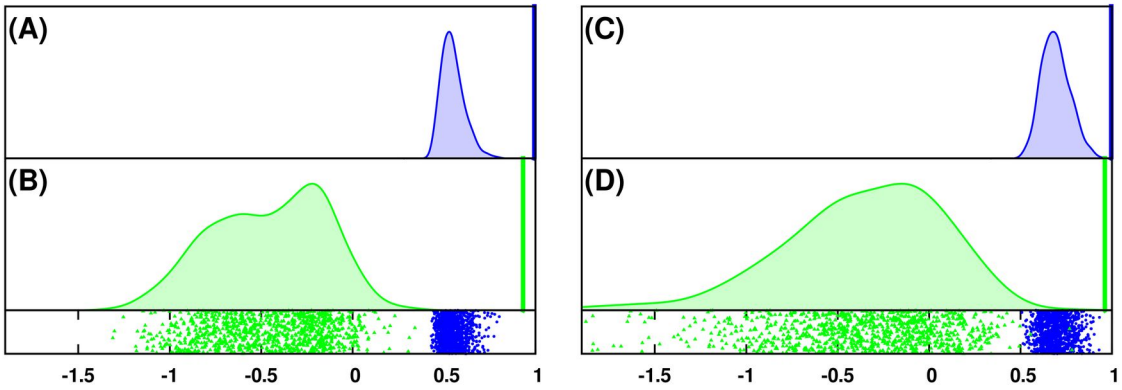

Figure 8.9: Response Permutation Test: GAI-vectorization.
Response permutation test results for OPLS-DA models from the GAI-binned (vectorized) liver (**A**, **B**) and fibroblast (**C**, **D**) data tensors. See the caption of Figure 8.6 for a complete description of the figure contents.

8.7 References

[1] K. M. Aberg, E. Alm, and R. J. Torgrip. The correspondence problem for metabonomics datasets. *Analytical and Bioanalytical Chemistry*, 394(1):151–162, 2009.

[2] P. E. Anderson, D. A. Mahle, T. E. Doom, N. V. Reo, N. J. DelRaso, and M. L. Raymer. Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics*, 7(2):179–190, 2011.

[3] P. E. Anderson, N. V. Reo, N. J. DelRaso, T. E. Doom, and M. L. Raymer. Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics*, 4(3):261–272, 2008.

[4] W. Astle, M. de Iorio, S. Richardson, D. Stephens, and T. Ebbels. A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures. *Journal of the American Statistical Association*, 107(500):37–41, 2012.

[5] R. Bro and A. K. Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

[6] O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson, and E. Holmes. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. *Analytical Chemistry*, 77(2):517–526, 2005.

[7] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson. Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, 85(1):144–154, 2007.

[8] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, 2000.

[9] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[10] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe – a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular NMR*, 6(3):277–293, 1995.

[11] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

[12] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[13] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[14] P. Eshghi. Dimensionality choice in principal components analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems*, 130:6–13, 2014.

[15] S. Halouska and R. Powers. Negative impact of noise on the principal component analysis of NMR data. *Journal of Magnetic Resonance*, 178(1):88–95, 2006.

[16] M. Hedenstrom, S. Wiklund, B. Sundberg, and U. Edlund. Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometrics and Intelligent Laboratory Systems*, 92(2):110–117, 2008.

[17] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[18] L. E. Kay, P. Keifer, and T. Saarinen. Pure Absorption Gradient Enhanced Heteronuclear Single Quantum Correlation Spectroscopy with Improved Sensitivity. *Journal of the American Chemical Society*, 114(26):10663–10665, 1992.

[19] W. J. Krzanowski. Cross-Validation in Principal Component Analysis. *Biometrics*, 43(3):575–584, 1987.

[20] J. C. Lindon, J. K. Nicholson, E. Holmes, and J. R. Everett. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance*, 12(5):289–320, 2000.

[21] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Uncorrelated Multilinear Principal Component Analysis for Unsupervised Multilinear Subspace Learning. *IEEE Transactions on Neural Networks*, 20(11):1820–1836, 2009.

[22] H. P. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, 2011.

[23] P. K. Mandal and A. Majumdar. A comprehensive discussion of HSQC and HMQC pulse sequences. *Concepts in Magnetic Resonance*, 20(1):1–23, 2004.

[24] A. Motta, D. Paris, and D. Melck. Monitoring Real-Time Metabolism of Living Cells by Fast Two-Dimensional NMR Spectroscopy. *Analytical Chemistry*, 82(6):2405–2411, 2010.

[25] A. G. Palmer, J. Cavanagh, P. E. Wright, and M. Rance. Sensitivity Improvement in Proton-Detected Two-Dimensional Heteronuclear Correlation NMR-Spectroscopy. *Journal of Magnetic Resonance*, 93(1):151–170, 1991.

[26] R. K. Rai and N. Sinha. Fast and Accurate Quantitative Metabolic Profiling of Body Fluids by Nonlinear Sampling of ^{1}H–^{13}C Two-Dimensional Nuclear Magnetic Resonance Spectroscopy. *Analytical Chemistry*, 84(22):10005–10011, 2012.

[27] S. A. A. Sousa, A. Magalhaes, and M. M. C. Ferreira. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122:93–102, 2013.

[28] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[29] Q. N. Van, H. J. Issaq, Q. Jiang, Q. Li, G. M. Muschik, T. J. Waybright, H. Lou, M. Dean, J. Uitto, and T. D. Veenstra. Comparison of 1D and 2D NMR spectroscopy for metabolic profiling. *Journal of Proteome Research*, 7(2):630–639, 2008.

[30] R. a. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(142):1–15, 2006.

[31] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[32] S. Wiklund, E. Johansson, L. Sjostrom, E. J. Mellerowicz, U. Edlund, J. P. Shockcor, J. Gottfries, T. Moritz, and J. Trygg. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, 80(1):115–122, 2008.

[33] B. Worley and R. Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

[34] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[35] Q. S. Xu, Y. Z. Liang, and Y. P. Du. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18(2):112–120, 2004.

[36] C. Zheng, S. C. Zhang, S. Ragg, D. Raftery, and O. Vitek. Identification and quantification of metabolites in ^{1}H NMR spectra by Bayesian model selection. *Bioinformatics*, 27(12):1637–1644, 2011.

Chapter 9
Multiblock Orthogonal Projections to Latent Structures

I see NIPALS as an open ended array of models with unlimited complexity in the combined use of several devices.

– Herman Wold

9.1 Introduction

The method of nonlinear iterative partial least squares (NIPALS) has firmly entrenched itself in the field of chemometrics. Implementations of principal component analysis (PCA) and projections to latent structures (PLS) that utilize NIPALS-type algorithms benefit from its numerical stability, as well as its flexibility and simplicity [1, 6, 21]. Only a few subroutines from level 2 of the basic linear algebra subprograms (BLAS) specification are required to construct a complete NIPALS-type algorithm [4], making it an attractive means of constructing PCA and PLS models of high-dimensional spectroscopic datasets.

One particularly recent addition to the NIPALS family of algorithms, called orthogonal projections to latent structures (OPLS), integrates an orthogonal signal correction (OSC) filter into NIPALS PLS [17, 2]. By extracting variation from its computed PLS components that is uncorrelated (orthogonal) to the responses, OPLS produces a more interpretable regression model compared to PLS. In fact, when trained on the same data and responses, and OPLS model and a PLS model with the same total number of components will show no difference in predictive ability [18]. Despite its relative novelty to the field, the enhanced interpretability of OPLS over PLS has made it a popular method in exploratory studies of spectroscopic datasets of complex chemical mixtures.

Extensions of NIPALS PCA and PLS to incorporate blocking information that partitions the set of measured variables into multiple "blocks" of data have recently gained attention in the field, as more experimental designs involve the collection of data from multiple analytical platforms per sample. In such experiments, referred to as "class II" multiblock schemes by Smilde et al. [15],

correlated consensus directions are sought from the blocks that maximally capture block variation and (optionally) maximally predict a set of responses. Of the available extensions of NIPALS to multiblock modeling, a class of methods exists that bears attractive computational qualities, namely computability from single-block bilinear factorizations. When both super weights and block loadings are normalized in consensus PCA (i.e. CPCA-W), the obtained super scores are equivalent to those obtained from PCA of the concatenated matrix of blocks [20]. Likewise, scores obtained from PLS of the concatenated matrix are equivalent to super scores from multiblock PLS (MB-PLS) when super scores are used in the deflation step [19, 20]. As a result, these multiblock bilinear factorizations inherit many of the useful properties of their single-block equivalents.

A second class of multiblock methods exists in which every block is predicted in a regression model by every other block. In the first of such methods, known as nPLS, the MAXDIFF criterion [16] is optimized one component at a time (i.e. sequentially) to yield a set of predictive weight vectors for each block [12]. The recently described OnPLS algorithm also falls within this class. OnPLS extends O2PLS to three or more matrices and may be considered a prefixing of nPLS with an OSC filtering step. OnPLS deflates non-globally predictive variation that may or may not be orthogonal to all blocks from each matrix, and then computes an nPLS model from the filtered result [12]. While fully symmetric OnPLS is a powerful and general addition to the existing set of multiblock modeling frameworks, it is arguably an over-complication when the regression of a single response matrix on multiple data blocks (i.e. MB-PLS) is sought. For such situations, a novel algorithm termed MB-OPLS for multiblock orthogonal projections to latent structures is introduced that embeds an OSC filter within NIPALS MB-PLS, thus solving an inherently different problem from OnPLS. It will be shown that MB-OPLS, in analogy to CPCA-W and MB-PLS, is computable from a single-block OPLS model of the matrix of concatenated data blocks. Thus, MB-OPLS forms a bridge between this special class of consensus component methods and the highly general symmetric regression framework of OnPLS.

9.2 Theory

MB-OPLS belongs to a set of multiblock methods that exhibit a computability from their single-block equivalents. A short discussion of these methods follows, in which the optimization criterion of each method is shown to belong to the MAXBET family of objective functions. This is contrasted to nPLS and OnPLS, which have been shown to optimize a MAXDIFF objective. Finally, the equiv-

alence of MB-OPLS and OPLS is demonstrated, and final mentions are made to differences between MB-OPLS and OnPLS.

In all following discussions, it will be understood that there exist n data matrices $\mathbf{X}_1$ to $\mathbf{X}_n$, each having N rows (observations) and K_i columns (variables). The matrix $\mathbf{X} = [\mathbf{X}_1 \mid \cdots \mid \mathbf{X}_n]$ of all concatenated blocks will be used in cases of single-block modeling. Finally, a response matrix $\mathbf{Y}$ having N rows and M columns will be assumed to exist for the purposes of regression.

9.2.1 nPLS and OnPLS

In their initial description of the OnPLS modeling framework [12], Löfstedt and Trygg introduced nPLS as a generalization of PLS regression to cases where $n > 2$, and a model is sought in which each matrix $\mathbf{X}_i$ is predicted by all other matrices $\mathbf{X}_{j \neq i}$. The nPLS solution involves identifying a set of weight vectors $\mathbf{w}_i$ that simultaneously maximize the covariances between each pair of resulting scores $\mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i$ via the following objective function:

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{n} {\mathbf{t}_i}^T \mathbf{t}_j = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} {\mathbf{w}_i}^T {\mathbf{X}_i}^T \mathbf{X}_j \mathbf{w}_j \tag{9.1}$$

subject to the constraints $\|\mathbf{w}_i\| = 1$. This objective was recognized to be a member of the MAXDIFF family of functions, whose solution is obtainable using a general algorithm from Hanafi and Kiers [5]. After the identification of a set of weight vectors, the scores

$$\mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i$$

and loadings

$$\mathbf{p}_i = \frac{{\mathbf{X}_i}^T \mathbf{t}_i}{{\mathbf{t}_i}^T \mathbf{t}_i}$$

may be computed for each matrix, which is then deflated prior to the computation of subsequent component weights:

$$\mathbf{X}_i \leftarrow \mathbf{X}_i - \mathbf{t}_i {\mathbf{p}_i}^T = \left(\mathbf{I} - \frac{\mathbf{t}_i {\mathbf{t}_i}^T}{{\mathbf{t}_i}^T \mathbf{t}_i} \right) \mathbf{X}_i \tag{9.2}$$

This deflation scheme follows the precedent set by two-block PLS regression. Because their described approach used a distinct deflation scheme from single-component (sequential) MAXDIFF, it was given the name "nPLS" by the authors to distinguish it from MAXDIFF [12, 8].

OnPLS extends nPLS by decomposing each matrix into a globally predictive part and a non-globally predictive (orthogonal) part using an orthogonal projection. By removing orthogonal variation from each block prior to constructing an nPLS model, OnPLS optimizes the following MAXDIFF-type objective function:

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbf{t}_i{}^T \mathbf{t}_j = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbf{w}_i{}^T \mathbf{X}_i{}^T \mathbf{Z}_i \mathbf{Z}_j \mathbf{X}_j \mathbf{w}_j \tag{9.3}$$

where $\mathbf{Z}_i$ represents the orthogonal projector identified by OnPLS for matrix i:

$$\mathbf{Z}_i = \mathbf{I} - \mathbf{T}_{\mathbf{o}i} \left(\mathbf{T}_{\mathbf{o}i}{}^T \mathbf{T}_{\mathbf{o}i} \right)^{-1} \mathbf{T}_{\mathbf{o}i}{}^T$$

where $\mathbf{T}_{\mathbf{o}i} = [\mathbf{t}_{\mathbf{o}i,1} \mid \cdots \mid \mathbf{t}_{\mathbf{o}i,A_o}]$, the concatenation of all orthogonal score vectors for the block, and $\mathbf{t}_{\mathbf{o}i,a} = \mathbf{X}_i \mathbf{w}_{\mathbf{o}i,a}$. In OnPLS, each orthogonal weight $\mathbf{w}_{\mathbf{o}i,a}$ is chosen such that its score $\mathbf{t}_{\mathbf{o}i,a}$ contains maximal covariance with the variation in $\mathbf{X}_{j \neq i}$ that is not jointly predictive of $\mathbf{X}_i$. The OnPLS framework provides a powerful set of methods for unsupervised data mining and path modeling [12, 9, 10, 11].

9.2.2 CPCA-W and MB-PLS

The consensus PCA method, introduced by Wold et al. as CPCA and modified by Westerhuis et al. into CPCA-W, identifies a set of weights $\mathbf{p}_i$ that maximally capture the within-block variances and between-block covariances of a set of n matrices [20]. It was further proven by Westerhuis, Kourti and MacGregor that the results of CPCA-W computed on matrices $\mathbf{X}_1$ to $\mathbf{X}_n$ are identical to those from PCA of the concatenated matrix $[\mathbf{X}_1 \mid \cdots \mid \mathbf{X}_n]$. It immediately follows from this equivalence that the CPCA-W algorithm optimizes the following objective function:

$$\mathbf{t}^T \mathbf{t} = \mathbf{p}^T \mathbf{X}^T \mathbf{X} \mathbf{p} = \sum_{i,j=1}^{n} \mathbf{t}_i{}^T \mathbf{t}_j = \sum_{i,j=1}^{n} \mathbf{p}_i{}^T \mathbf{X}_i{}^T \mathbf{X}_j \mathbf{p}_j \tag{9.4}$$

subject to the constraint $\|\mathbf{p}\| = 1$, where $\mathbf{p}^T = [\mathbf{p}_1{}^T \mid \cdots \mid \mathbf{p}_n{}^T]$. Maximizing the above function yields a set of super scores $\mathbf{t}$ that relate the N observations in $\mathbf{X}$ to each other based on the extracted consensus in $\mathbf{p}$, as well as block scores $\mathbf{t}_i$ and loadings $\mathbf{p}_i$ that describe each block. This objective function is of the MAXBET variety, in contrast to the MAXDIFF objective of nPLS and OnPLS. As a result, the CPCA-W NIPALS algorithm may be considered a special case of the general algorithm

from Hanafi and Kiers [5].

The multiblock PLS (MB-PLS) method, when deflation is performed using super scores [19], shares an equivalence with single-block PLS as proven by Westerhuis et al. [20]. Therefore, the MB-PLS objective takes on a similar form as in CPCA-W, with the addition of a weighting matrix:

$$\mathbf{t}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \sum_{i,j=1}^{n} {\mathbf{t}_i}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t}_j = \sum_{i,j=1}^{n} {\mathbf{w}_i}^T{\mathbf{X}_i}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j\mathbf{w}_j \tag{9.5}$$

where once again $\|\mathbf{w}\|$ is constrained to unity. In analogy to Höskuldsson's interpretation of PLS as a regression on orthogonal components, where $\mathbf{Y}\mathbf{Y}^T$ is used to weight the covariance matrix, the above function corresponds to a MAXBET objective with an inner weighting of $\mathbf{Y}\mathbf{Y}^T$ [6]. Alternatively, equation 9.5 could be interpreted as a MAXBET computed on the n cross-covariance matrices $\mathbf{Y}^T\mathbf{X}_1$ to $\mathbf{Y}^T\mathbf{X}_n$.

9.2.3 MB-OPLS

Extension of prior multiblock NIPALS algorithms to incorporate an OSC filter rests on the observation that, in both the cases of CPCA-W and MB-PLS, deflation of each computed component is accomplished using super scores. For any super score deflation method, a loading vector is computed for each block:

$$\mathbf{p}_i = \frac{{\mathbf{X}_i}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$$

and the super scores $\mathbf{t}$ and block loadings are then used to deflate their respective block:

$$\mathbf{X}_i \leftarrow \mathbf{X}_i - \mathbf{t}{\mathbf{p}_i}^T = \left(\mathbf{I} - \frac{\mathbf{t}\mathbf{t}^T}{\mathbf{t}^T\mathbf{t}}\right)\mathbf{X}_i \tag{9.6}$$

Equation 9.6 differs from equation 9.2 used in nPLS and OnPLS, which uses block-specific scores and loadings during deflation. This method of super score deflation ensures that the super scores become an orthogonal basis, while allowing scores and loadings to become slightly correlated at the block level, and is a necessary condition for the equivalences between CPCA-W and MB-OPLS and their single-block counterparts [20]. This condition shall be employed in MB-OPLS by deflating each matrix by a set of orthogonal super scores $\mathbf{T_o}$, which shall be shown to be equal to the orthogonal scores obtained from single-block OPLS. By constructing an MB-PLS model on the set of matrices

after deflation by $\mathbf{T_o}$, we effectively arrive at another MAXBET objective:

$$\mathbf{t}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \mathbf{w}^T\mathbf{X}^T\mathbf{Z}\mathbf{Y}\mathbf{Y}^T\mathbf{Z}\mathbf{X}\mathbf{w} = \sum_{i,j=1}^{n} \mathbf{t}_i{}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t}_j = \sum_{i,j=1}^{n} \mathbf{w}_i{}^T\mathbf{X}_i{}^T\mathbf{Z}\mathbf{Y}\mathbf{Y}^T\mathbf{Z}\mathbf{X}_j\mathbf{w}_j \tag{9.7}$$

where $\mathbf{w}$ is constrained to unit ℓ_2-norm and $\mathbf{Z}$ is the orthogonal projector for the super scores $\mathbf{T_o}$.

The MB-OPLS Model

MB-OPLS constructs an OPLS model for each matrix $\mathbf{X}_i$, where the predictive and orthogonal loadings for each matrix are interrelated by a set of predictive and orthogonal super scores, respectively:

$$\mathbf{X}_i = \underbrace{\mathbf{T}\mathbf{P}_i{}^T}_{\mathbf{X_p}_i} + \underbrace{\mathbf{T_o}\mathbf{P_o}_i{}^T}_{\mathbf{X_o}_i} + \mathbf{E}_i \tag{9.8}$$

Concatenation of all block-level matrices together in equation 9.8 results in a top-level consensus model, which is in fact equivalent to an OPLS model trained on the partitioned data matrix $\mathbf{X}$:

$$\mathbf{X} = [\mathbf{X}_1 \mid \cdots \mid \mathbf{X}_n] = \underbrace{\mathbf{T}[\mathbf{P}_1{}^T \mid \cdots \mid \mathbf{P}_n{}^T]}_{\mathbf{X_p}} + \underbrace{\mathbf{T_o}[\mathbf{P_o}_1{}^T \mid \cdots \mid \mathbf{P_o}_n{}^T]}_{\mathbf{X_o}} + \underbrace{[\mathbf{E}_1 \mid \cdots \mid \mathbf{E}_n]}_{\mathbf{E}} \tag{9.9}$$

Like PLS, OPLS and MB-PLS, an MB-OPLS model contains a second equation that relates the predictive super scores and responses:

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^T + \mathbf{F} \tag{9.10}$$

The MB-OPLS Algorithm

The MB-OPLS algorithm was introduced in Chapter 3, but will be decomposed in more detail as sub-algorithms here in order to relate it with the OPLS NIPALS algorithm. For simplicity, procedures relating to cross-validation and automatic component count identification will be stripped from this discussion of MB-OPLS. MB-OPLS admits a matrix of responses $\mathbf{Y}$, but also supports vector-$\mathbf{y}$ modeling as a special case. Direct and normed assignment will be indicated by "$\leftarrow$" and "$\propto$", respectively.

Algorithm 9.1 Core Algorithm for MB-OPLS

Input: $\{\mathbf{X}_i \in \mathbb{R}^{N \times K_i}\}_{i=1}^{n}, \mathbf{Y} \in \mathbb{R}^{N \times M}$

1: $\mathbf{V}_i \leftarrow \text{SUBSPACE}(\mathbf{X}_i, \mathbf{Y}) \quad \forall i \in \{1, \ldots, n\}$
2: $\{\mathbf{w}_i\}_{i=1}^{n}, \{\mathbf{t}_i\}_{i=1}^{n}, \{\mathbf{p}_i\}_{i=1}^{n}, \mathbf{t}, \mathbf{c}, \mathbf{u} \leftarrow \text{PREDCMP}(\{\mathbf{X}_i\}_{i=1}^{n}, \mathbf{Y})$
3: To compute an orthogonal component, continue to step (4).
   Otherwise, proceed to step (7).
4: $\{\mathbf{w}_{\mathbf{o}i}\}_{i=1}^{n}, \{\mathbf{t}_{\mathbf{o}i}\}_{i=1}^{n}, \{\mathbf{p}_{\mathbf{o}i}\}_{i=1}^{n}, \mathbf{t}_{\mathbf{o}} \leftarrow \text{ORTHCMP}(\{\mathbf{X}_i\}_{i=1}^{n}, \{\mathbf{V}_i\}_{i=1}^{n}, \{\mathbf{p}_i\}_{i=1}^{n})$
5: $\mathbf{X}_i \leftarrow \mathbf{X}_i - \mathbf{t}_{\mathbf{o}}{\mathbf{p}_{\mathbf{o}}}_i^T \quad \forall i \in \{1, \ldots, n\}$
6: Return to step (2).
7: $\mathbf{X}_i \leftarrow \mathbf{X}_i - \mathbf{t}\mathbf{p}_i^T \quad \forall i \in \{1, \ldots, n\}$
8: To compute another predictive component, return to step (2).
   Otherwise, end.

The SUBSPACE method computes the $\mathbf{Y}$-predictive subspace for each data matrix $\mathbf{X}_i$, and follows directly from the single-block OPLS algorithm described by Trygg and Wold [17]:

Algorithm 9.2 Predictive Subspace Identification for MB-OPLS

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}, \mathbf{Y} \in \mathbb{R}^{N \times M}$

1: **for all** $m \in \{1, \ldots, M\}$ **do**
2: $\quad \mathbf{v}_m \leftarrow \mathbf{X}^T\mathbf{y}_m \cdot \left(\mathbf{y}_m^T\mathbf{y}_m\right)^{-1}$
3: $\quad \mathbf{V} \leftarrow [\mathbf{V} \mid \mathbf{v}_m]$
4: **end for**

Predictive MB-OPLS components identified by the PREDCMP method are, in fact, MB-PLS components:

Algorithm 9.3 Predictive Component Computation for MB-OPLS

Input: $\{\mathbf{X}_i \in \mathbb{R}^{N \times K_i}\}_{i=1}^{n}$, $\mathbf{Y} \in \mathbb{R}^{N \times M}$

1: Initialize $\mathbf{u}$ to a column of $\mathbf{Y}$
2: **repeat**
3: $\quad \mathbf{w}_i \propto {\mathbf{X}_i}^T \mathbf{u} \quad \forall i \in \{1, \ldots, n\}$
4: $\quad \mathbf{t}_i \leftarrow \mathbf{X}_i \mathbf{w}_i \quad \forall i \in \{1, \ldots, n\}$
5: $\quad \mathbf{R} \leftarrow [\mathbf{t}_1 \mid \cdots \mid \mathbf{t}_n]$
6: $\quad \mathbf{w_T} \propto \mathbf{R}^T \mathbf{u}$
7: $\quad \mathbf{t} \leftarrow \mathbf{R}\mathbf{w_T}$
8: $\quad \mathbf{c} \leftarrow \left(\mathbf{Y}^T \mathbf{t}\right) \cdot \left(\mathbf{t}^T \mathbf{t}\right)^{-1}$
9: $\quad \mathbf{u} \leftarrow (\mathbf{Y}\mathbf{c}) \cdot \left(\mathbf{c}^T \mathbf{c}\right)^{-1}$
10: **until** $\|\mathbf{u} - \mathbf{u}_{old}\| \cdot \|\mathbf{u}_{old}\|^{-1} < \varepsilon$
11: $\mathbf{p}_i \leftarrow \left({\mathbf{X}_i}^T \mathbf{t}\right) \cdot \left(\mathbf{t}^T \mathbf{t}\right)^{-1} \quad \forall i \in \{1, \ldots, n\}$

In the above method, the value of ε is set to a very small number, such as 10^{-9}. Once a predictive component has been computed, MB-OPLS uses the ORTHCMP method to extract a new orthogonal component:

Algorithm 9.4 Orthogonal Component Computation for MB-OPLS

Input: $\{\mathbf{X}_i \in \mathbb{R}^{N \times K_i}\}_{i=1}^{n}$, $\{\mathbf{V}_i \in \mathbb{R}^{M \times K_i}\}_{i=1}^{n}$, $\{\mathbf{p}_i \in \mathbb{R}^{K_i}\}_{i=1}^{n}$

1: $\mathbf{w_o}_i \leftarrow \mathbf{p}_i \quad \forall i \in \{1, \ldots, n\}$
2: **for all** $m \in \{1, \ldots, M\}$ **do**
3: $\quad \phi \leftarrow \left(\sum_{i=1}^{n} {\mathbf{v}_{i,m}}^T \mathbf{w_o}_i\right) \cdot \left(\sum_{i=1}^{n} {\mathbf{v}_{i,m}}^T \mathbf{v}_{i,m}\right)^{-1}$
4: $\quad \mathbf{w_o}_i \leftarrow \mathbf{w_o}_i - \phi \mathbf{v}_{i,m} \quad \forall i \in \{1, \ldots, n\}$
5: **end for**
6: $\alpha \leftarrow \left(\sum_{i=1}^{n} {\mathbf{w_o}_i}^T \mathbf{w_o}_i\right)^{-1/2}$
7: $\mathbf{w_o}_i \leftarrow \alpha \mathbf{w_o}_i \quad \forall i \in \{1, \ldots, n\}$
8: $\mathbf{t_o}_i \leftarrow \mathbf{X}_i \mathbf{w_o}_i \quad \forall i \in \{1, \ldots, n\}$
9: $\mathbf{t_o} \leftarrow \sum_{i=1}^{n} \mathbf{t_o}_i$
10: $\mathbf{p_o}_i \leftarrow \left({\mathbf{X}_i}^T \mathbf{t_o}\right) \cdot \left({\mathbf{t_o}}^T \mathbf{t_o}\right)^{-1} \quad \forall i \in \{1, \ldots, n\}$

For each predictive component in the model, a set of orthogonal components is extracted. After the computation of a new orthogonal component, the current predictive component is updated to reflect the removal of orthogonal variation from the matrices $\mathbf{X}_i$. The MB-OPLS algorithm closely follows the matrix-$\mathbf{Y}$ OPLS algorithm presented by Trygg and Wold [17], but replaces the standard PLS computation (steps 4–10 in OPLS) with an MB-PLS computation.

Equivalence to OPLS

In both the vector-$\mathbf{y}$ and matrix-$\mathbf{Y}$ OPLS algorithms proposed by Trygg and Wold [17], a basis $\mathbf{V}$ for the response-correlated variation in $\mathbf{X}$ is constructed by regressing the data onto each column of

responses:

$$\mathbf{v}_m \leftarrow \frac{\mathbf{X}^T\mathbf{y}_m}{{\mathbf{y}_m}^T\mathbf{y}_m} \quad \forall m \in \{1,\ldots,M\} \tag{9.11}$$

where $\mathbf{y}_m$ and $\mathbf{v}_m$ denote the m-th columns of $\mathbf{Y}$ and $\mathbf{V}$, respectively. When $\mathbf{X}$ is partitioned into multiple blocks, the computed basis also bears the same partitioning, i.e. $\mathbf{V}^T = [{\mathbf{V}_1}^T \mid \cdots \mid {\mathbf{V}_n}^T]$, where each of the n submatrices corresponds to the regression of its respective block $\mathbf{X}_i$ onto the responses:

$$\mathbf{v}_{i,m} \leftarrow \frac{{\mathbf{X}_i}^T\mathbf{y}_m}{{\mathbf{y}_m}^T\mathbf{y}_m} \quad \forall m \in \{1,\ldots,M\} \tag{9.12}$$

where $\mathbf{v}_{i,m}$ is the m-th column of $\mathbf{V}_i$. Therefore, the bases of response-correlated variation identified by OPLS and MB-OPLS are equal. Given a single-block PLS loading vector $\mathbf{p}$, the OPLS algorithm computes an orthogonal weight $\mathbf{w_o}$ by orthogonalizing $\mathbf{p}$ to the columns of $\mathbf{V}$:

$$\mathbf{w_o} \leftarrow \mathbf{w_o} - \left(\frac{{\mathbf{v}_m}^T\mathbf{w_o}}{{\mathbf{v}_m}^T\mathbf{v}_m}\right)\cdot\mathbf{v}_m \quad \forall m \in \{1,\ldots,M\} \tag{9.13}$$

after $\mathbf{w_o}$ has been initialized from $\mathbf{p}$. From the proof of Westerhuis et al. [20], it is known that the single-block PLS loading $\mathbf{p}$ equals the concatenation of all block loadings from MB-PLS, i.e. that $\mathbf{p}^T = [{\mathbf{p}_1}^T \mid \cdots \mid {\mathbf{p}_n}^T]$. Expansion of all vector terms in the above equation into their partitioned forms results in the following new assignment rule:

$$\mathbf{w_o}_i \leftarrow \mathbf{w_o}_i - \left(\frac{\sum_{i=1}^{n}{\mathbf{v}_{i,m}}^T\mathbf{w_o}_i}{\sum_{i=1}^{n}{\mathbf{v}_{i,m}}^T\mathbf{v}_{i,m}}\right)\cdot\mathbf{v}_{i,m} \quad \forall m \in \{1,\ldots,M\} \tag{9.14}$$

The scalar term in equation 9.14 should be recognized as ϕ in the ORTHCMP method. By the same reasoning, steps 6 and 7 in ORTHCMP are equivalent to scaling $\mathbf{w_o}$ to unit norm. Therefore, because $\mathbf{w_o}$ is the column-wise concatenation of all weights $\mathbf{w_o}_i$, it is then apparent that the orthogonal super scores extracted by MB-OPLS are identical to those from OPLS of the concatenated matrix $\mathbf{X}$:

$$\mathbf{t_o} = \mathbf{X}\mathbf{w_o} = [\mathbf{X}_1 \mid \cdots \mid \mathbf{X}_n]\begin{bmatrix}\mathbf{w_o}_1\\ \vdots \\ \mathbf{w_o}_n\end{bmatrix} = \sum_{i=1}^{n}\mathbf{X}_i\mathbf{w_o}_i = \sum_{i=1}^{n}\mathbf{t_o}_i \tag{9.15}$$

From this equivalence, and the fact that PREDCMP in MB-OPLS constitutes an MB-PLS computation, we arrive at the equivalence between MB-OPLS and OPLS. Thus, orthogonality between the responses and orthogonal super scores $\mathbf{t_o}$ computed by MB-OPLS is also ensured. However, because the computation of orthogonal weights involves all blocks, the resulting orthogonal block scores $\mathbf{t_o}_i$

are not guaranteed to be orthogonal to the responses.

Computation from an OPLS Model

The equivalence between MB-OPLS super scores and OPLS scores may be leveraged to generate an MB-OPLS model from an existing OPLS model of a partitioned data matrix, saving computation time during cross-validated model training. Algorithm 3.7 in Chapter 3 details the extraction of MB-OPLS block scores and loadings from an OPLS model.

The keen observer will recognize the similarity between Algorithm 3.7 and the procedure outlined by Westerhuis et al. for extracting MB-PLS block components from a PLS model [20]. By using this algorithm to compute MB-OPLS models, the analyst avoid the unnecessary computation of block components during cross-validated model training.

9.3 Datasets

Two datasets will be described to illustrate how MB-OPLS effectively integrates an OSC filter into an MB-PLS decomposition of a set of n matrices. The first synthetic dataset contrasts the mixing of predictive information in MB-PLS with its separation in MB-OPLS using a contrived three-block regression example similar to that introduced by Löfstedt and Trygg [12]. The second dataset, a joint set of NMR and MS observations introduced in Chapter 4, is used to demonstrate the enhanced interpretability of MB-OPLS models over MB-PLS in a real example of discriminant analysis. All modeling and validation were performed using routines available in the MVAPACK chemometrics toolbox [22].

9.3.1 Synthetic Example

In the first dataset, three matrices (all having 100 rows and 200 columns) were constructed to hold one $\mathbf{y}$-predictive component ($\mathbf{tp}_i{}^T$) and one $\mathbf{y}$-orthogonal component ($\mathbf{t_o p_o}_i{}^T$). The score vectors were non-overlapping (orthogonal) Gaussian density functions, and all block loading vectors were mutually overlapping Gaussian density or square step functions. The true synthetic block loadings are illustrated in Figure 9.1A. A two-component MB-PLS-R model was trained on the synthetic three-block example dataset, as well as a $1 + 1$ (one predictive, one orthogonal) component MB-OPLS-R model. Block loadings extracted by MB-PLS-R and MB-OPLS-R and shown in Figures

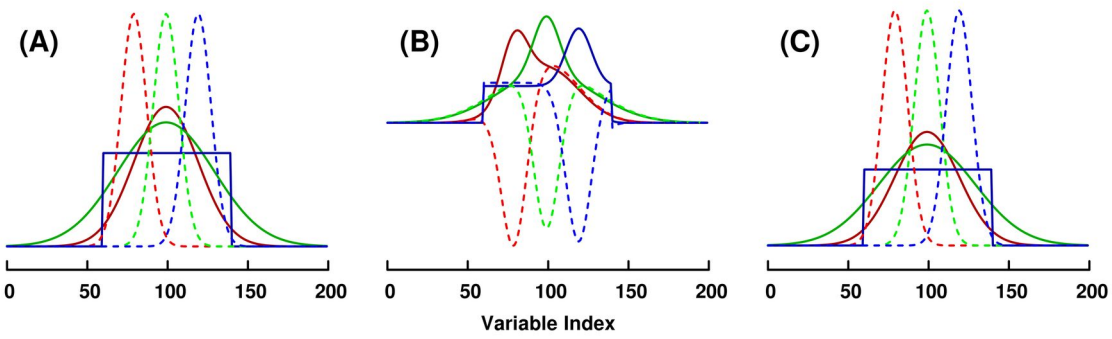

Figure 9.1: Synthetic Three-block Example Dataset.
Block loadings in the synthetic multiblock example dataset. (**A**) True predictive loadings (solid) and orthogonal loadings (dashed) used to construct the three-block dataset. First, second and third block loadings are colored in red, green and blue, respectively. (**B**) First component (solid) and second component (dashed) loadings identified by MB-PLS modeling of the three data blocks. (**C**) Predictive (solid) and orthogonal (dashed) block loadings identified by MB-OPLS, illustrating the separation of **y**-uncorrelated variation accomplished by the integrated OSC filter.

9.1B and 9.1C, respectively.

9.3.2 Joint ^{1}H NMR and DI-ESI-MS Datasets

The second dataset is a pair of processed and treated data matrices, collected on 29 samples of metabolite extracts from human dopaminergic neuroblastoma cells treated with various neurotoxic agents [7]. Details about the collection, processing and treatment of this dataset may be found in Chapter 4, but will be summarized here. The first matrix, collected using ^{1}H NMR spectroscopy, contains 16,138 columns and the second, collected using direct injection electrospray ionization mass spectrometry (DI-ESI-MS), contains 2,095 columns. Prior to all modeling, block weighting was applied after Pareto scaling to ensure equal contribution of each block to the models [15].

In a previously published analysis of this dataset [13], a two-component, two-class (vector-**y**) MB-PLS-DA model was trained on the dataset in order to discriminate between untreated and neurotoxin-treated cell samples. To highlight the improved interpretability of MB-OPLS over MB-PLS, a $1+1$ MB-OPLS-DA model was trained on the data using an identical vector of class labels. Block components were extracted from an OPLS-DA model of the concatenated matrix $\mathbf{X} = [\mathbf{X}_{NMR} \mid \mathbf{X}_{MS}]$ using Algorithm 3.7 in Chapter 3. For both models, fifty rounds of Monte Carlo seven-fold cross-validation [14, 23] were performed to compute per-component Q^2 statistics [21], in addition to the R^2 statistics available from model training. CV-ANOVA significance testing was also applied to further assess model reliability [3].

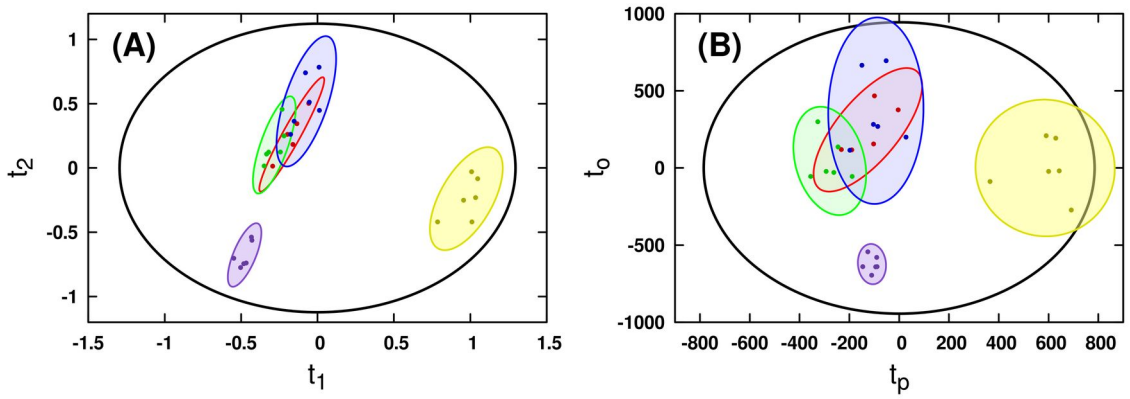

Figure 9.2: Super Scores of Joint Spectroscopic Data.
Super scores identified by (**A**) MB-PLS and (**B**) MB-OPLS modeling of the joint ^{1}H NMR and DI-ESI-MS data matrices. Extraction of **y**-orthogonal variation from the first PLS component is clear in the MB-OPLS scores. Ellipses represent the 95% confidence regions for each sub-class of observations, assuming normal distributions. Colors indicate membership to the untreated (yellow), 6-hydroxydopamine (red), 1-methyl-4-phenylpyridinium (green) and paraquat (violet) sub-classes. Cross-validated super scores are shown in Figure 9.3.

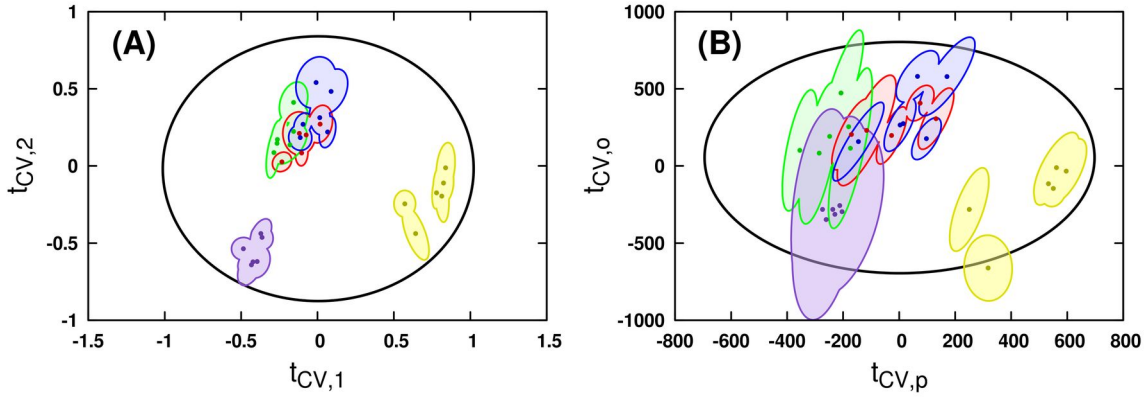

Figure 9.3: Cross-validated Super Scores of Joint Spectroscopic Data.
Cross-validation estimated super scores from (**A**) MB-PLS and (**B**) MB-OPLS modeling of the joint ^{1}H NMR and DI-ESI-MS data matrices. Points indicate mean values for each observation, and filled regions represent the union of all observations' confidence intervals from Monte Carlo iterations. Class colors are identical to those in Figure 9.2.

9.4 Results and Discussion

In both the contrived dataset and the real spectroscopic dataset, the interpretative advantage offered by MB-OPLS over MB-PLS is strikingly apparent. In the synthetic example, MB-OPLS capably identifies the true predictive and orthogonal loadings in the presence of **y**-orthogonal variation that clouds the interpretation of MB-PLS loadings (Figure 9.1). By design, this comparison between MB-OPLS and MB-PLS is highly similar to the first example presented by Löfstedt and Trygg to

compare nPLS and OnPLS for general data discovery [12]. However, as is evidenced by the differences between equations 9.3 and 9.7 above, MB-OPLS solves an inherently distinct problem from OnPLS: the identification of consensus variation in multiple blocks of data that predicts a single set of responses.

The ability of MB-OPLS to separate predictive and orthogonal variation from multiple data matrices is further exemplified in the discriminant analysis of the real spectroscopic dataset. From the rotated discrimination axis in the MB-PLS-DA scores (Figure 9.2A), it is clear that predictive and orthogonal variation have become mixed in the corresponding block loadings (Figure 4.11). Integration of an OSC filter into the multiblock model in the form of MB-OPLS-DA achieved the expected rotation of super scores to place more predictive variation into the first component (Figure 9.2B). As a consequence of this rotation, spectral information that separates paraquat treatment from other neurotoxin treatments is also moved into the orthogonal component. For example, strong loadings from citrate in the ^{1}H NMR MB-PLS block loadings (Figure 4.11A, 2.6 ppm) are substantially diminished in the predictive block loadings from MB-OPLS (Figure 4.12A), as separation between paraquat and other treatments has been isolated along the orthogonal component in super scores. Inspection of the orthogonal block loadings from MB-OPLS (Figure 4.13) will reveal, as expected, that citrate contributes more to separation between neurotoxin treatments than to separation between treatments and controls.

The partial correlation of both predictive and orthogonal block scores in MB-OPLS is readily observed in the comparison of block scores from MB-PLS and MB-OPLS (Figures 9.4 and 9.5). While the super scores in Figure 9.2B are rotated to separate predictive and orthogonal variation, block scores in Figures 9.4B and 9.5B have slightly rotated back into alignment with the MB-PLS block scores. This partial correlation and re-mixing of predictive and orthogonal variation in MB-OPLS block scores is a consequence of the use of super score deflation in the presented algorithm. When all matrices contain similar patterns of orthogonal variation, their MB-OPLS block scores will reflect this by retaining the OSC-induced rotation captured at the consensus level by the super scores.

Because the MB-OPLS-DA model of the real spectral data matrices was trained using the single-block OPLS routine already present in MVAPACK, all standard cross-validation metrics were available in the model without further computational expenditure. Monte Carlo cross-validation of the MB-PLS model produced cumulative R^2_Y and Q^2 statistics of 0.988 and 0.901$\pm$0.015, respectively,

and validation of the MB-OPLS model resulted in statistics of 0.903 and 0.706±0.028, respectively. In addition, MB-OPLS modeling yielded $R^2_{X,p}$ and $R^2_{X,o}$ statistics of 0.378 and 0.245 for the first block, and 0.236 and 0.083 for the second block. Monte Carlo cross-validated super scores from MB-PLS and MB-OPLS are depicted in Figure 9.2. Compared to MB-PLS scores in Figure 9.2A, MB-OPLS scores (Figure 9.2B) exhibit an increased uncertainty due to the coupled nature of predictive and orthogonal components in OPLS models. Further validation of the MB-OPLS-DA model via CV-ANOVA produced a p value equal to 2.88×10^{-6}, indicating a sufficiently reliable model.

It is worthy of final mention that the objective solved by MB-OPLS is but a single member of a super-family of methods introduced in detail by Hanafi and Kiers [5]. In the first family, nPLS and OPLS maximally capture the between-matrix covariances before and after orthogonal signal correction, respectively, and thus serve to regress a set of matrices against each other. Methods in the second family capture *both* within-matrix variances and between-matrix covariances of a set of matrices (CPCA-W), a set of response-weighted matrices (MB-PLS), and a set of response-weighted OSC-filtered matrices (MB-OPLS). By casting these methods in the light of MAXDIFF and MAXBET, we obtain an informative picture of their characteristics, commonalities, and differences. For example, nPLS and OnPLS force an equal contribution of each matrix to the solution through the constraint $\|\mathbf{w}_i\| = 1$, while CPCA-W, MB-PLS and MB-OPLS allow contributions to float based on the "importance" of each matrix to the modeling problem at hand. This super weight approach necessitates a block scaling procedure to avoid highly weighting any given matrix due to size alone [15, 20].

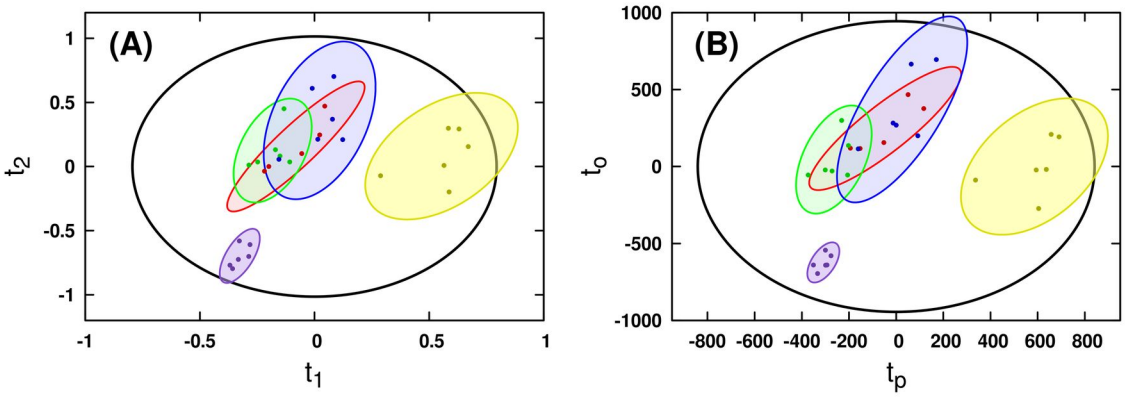

Figure 9.4: First-block Scores of Joint Spectroscopic Data.
Block scores from (**A**) MB-PLS and (**B**) MB-OPLS modeling of the ^{1}H NMR data matrix. Ellipses and class colors are identical to those in Figure 9.2.

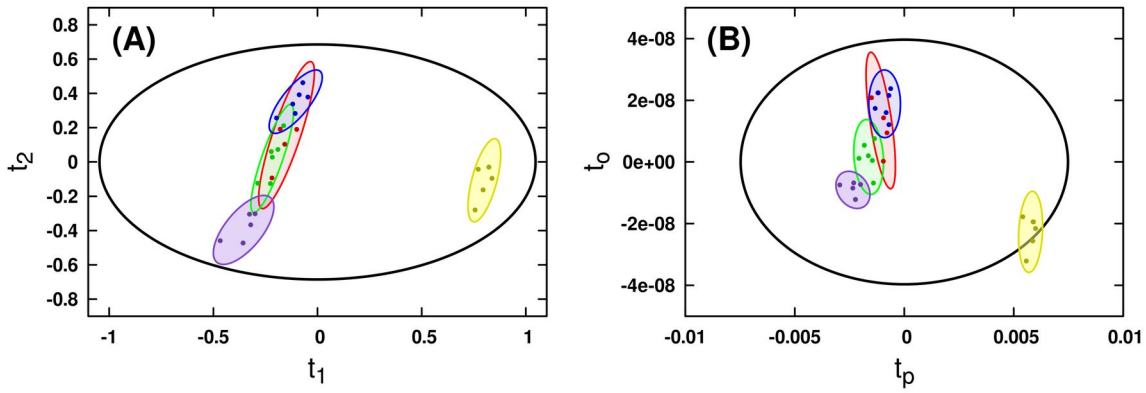

Figure 9.5: Second-block Scores of Joint Spectroscopic Data.
Block scores from (**A**) MB-PLS and (**B**) MB-OPLS modeling of the DI-ESI-MS data matrix. Ellipses and class colors are identical to those in Figure 9.2.

9.5 Conclusions

The MB-OPLS method described in this chapter is a versatile extension of MB-PLS to include an OSC filter, and belongs to a family of MAXBET optimizers that share an equivalence with their single-block factorizations. By removing consensus response-uncorrelated variation from a set of n data matrices, MB-OPLS expands the scope and benefits of OPLS to cases where blocking information is available. The ability of MB-OPLS to separate predictive and orthogonal variation from multiple blocks of data has been demonstrated on both synthetic and real spectral data, both in cases of vector-$\mathbf{y}$ regression and discriminant analysis. The described algorithm admits either a vector or a matrix as responses, and is implemented in the latest version of the open-source MVAPACK chemometrics toolbox [22].

9.6 References

[1] M. Andersson. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23:518–529, 2009.

[2] J.-C. Boulet and J.-M. Roger. Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61–69, 2012.

[3] L. Eriksson, J. Trygg, and S. Wold. CV-ANOVA for significance testing of PLS and OPLS models. *Journal of Chemometrics*, 22(11-12):594–600, 2008.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 4 edition, 2012.

[5] M. Hanafi and H. A. L. Kiers. Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Computational Statistics and Data Analysis*, 51:1491–1508, 2006.

[6] A. Hoeskuldsson. PLS Regression Methods. *Journal of Chemometrics*, 2:211–228, 1988.

[7] S. Lei, L. Zavala-Flores, A. Garcia-Garcia, R. Nandakumar, Y. Huang, N. Madayiputhiya, R. C. Stanton, E. D. Dodds, R. Powers, and R. Franco. Alterations in Energy/Redox Metabolism Induced by Mitochondrial and Environmental Toxins: A Specific Role for Glucose-6-Phosphate Dehydrogenase and the Pentose Phosphate Pathway in Paraquat Toxicity. *ACS Chemical Biology*, 9(9):2032–2048, 2014.

[8] T. Lofstedt. *OnPLS: Orthogonal Projections to Latent Structures in Multiblock and Path Model Data Analysis*. PhD thesis, Umea University, 2012.

[9] T. Lofstedt, L. Eriksson, G. Wormbs, and J. Trygg. Bi-modal OnPLS. *Journal of Chemometrics*, 26:236–245, 2012.

[10] T. Lofstedt, M. Hanafi, G. Mazerolles, and J. Trygg. OnPLS path modelling. *Chemometrics and Intelligent Laboratory Systems*, 118:139–149, 2012.

[11] T. Lofstedt, D. Hoffman, and J. Trygg. Global, local and unique decompositions in OnPLS for multiblock data analysis. *Analytica Chimica Acta*, 791:13–24, 2013.

[12] T. Lofstedt and J. Trygg. OnPLS – a novel multiblock method for the modeling of predictive and orthogonal variation. *Journal of Chemometrics*, 25:441–455, 2011.

[13] D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015.

[14] J. Shao. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[15] A. K. Smilde, J. A. Westerhuis, and S. de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337, 2003.

[16] J. M. F. ten Berge, H. A. L. Kiers, and J. de Leeuw. Explicit CANDECOMP/PARAFAC solutions for a contrived 2x2x2 array of rank three. *Psychometrika*, 53(4):579–583, 1988.

[17] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[18] T. Verron, R. Sabatier, and R. Joffre. Some theoretical properties of the O-PLS method. *Journal of Chemometrics*, 18:62–68, 2004.

[19] J. A. Westerhuis and P. M. J. Coenegracht. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11(5):379–392, 1997.

[20] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[21] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[22] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[23] Q. S. Xu and Y. Z. Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

Chapter 10
Quantification of PCA/PLS-DA Class Separations

People want to see patterns in the world. ... So important is this skill that we apply it everywhere, warranted or not.

– Benoit Mandelbrot

10.1 Introduction

The importance placed on interpretation of PCA, PLS-DA and OPLS-DA scores plots necessitates the use of quantitative procedures to determine the significance of separations between multiple experimental groups in scores space. However, no de facto protocol or metric exists to provide a means of reporting the degree or significance of group separation [18, 9, 8]. Anderson et al. used the J_2 criterion [1, 12] to assess the quality of resulting scores clusters according to the average withing-group and between-group scatters for all groups. However, the J_2 metric only provides an overall estimate of group separation without fine-grained information on each pair of groups [12]. A similar problem exists with the related Davies-Bouldin index [2], which chooses a worst-case estimate of group overlap as its figure of merit. Dixon et al. [4] also comprehensively reported the performances of four cluster separation indices based on modifications of metrics used to validate separation for unsupervised clustering algorithms. Alternatively, the PCAtoTree protocol constructs dendrograms from Euclidean distance matrices computed from PCA scores for the PHYLIP [6] software suite using a bootstrapping routine to determine branch node significance [18, 16]. However, it was recently shown that hypothesis testing using a Mahalanobis distance metric and the T^2 and F distributions can provide a statistical means of quantifying group similarity [8], suggesting the possibility of returning p values for full statistical quantitation of group separations in scores space.

10.2 Materials and Methods

The methods described below were implemented in software using the C programming language with minimal external dependencies, so the programs may be compiled and executed on any modern

GNU/Linux distribution.

10.2.1 Probability Calculation

Under the assumption that each group in the scores space is distributed as a multivariate normal random variable, the separations between groups may be calculated using the squared Mahalanobis distance metric [14]:

$$D_M^2 = (\mathbf{u}_j - \mathbf{u}_i)^T \mathbf{S}_p^{-1} (\mathbf{u}_j - \mathbf{u}_i) \tag{10.1}$$

In the above equation, $\mathbf{u}_i, \mathbf{u}_j \in \mathbb{R}^p$ are the p-variate sample means of groups i and j, respectively, and $\mathbf{S}_p \in \mathbb{R}^{p \times p}$ is the pooled variance-covariance matrix, a weighted sum of the covariance matrices from groups i and j:

$$\mathbf{S}_p = \frac{n_i \mathbf{S}_i + n_j \mathbf{S}_j}{n_i + n_j} \tag{10.2}$$

where n_i and n_j are the number of observations in groups i and j, respectively. The Mahalanobis distance may then be related to a Hotelling's T^2 statistic by the following scaling [15]:

$$T^2 = \left(\frac{n_i n_j}{n_i + n_j} \right) D_M^2 \tag{10.3}$$

This T^2 statistic is an extension of the Student's t statistic to hypothesis tests in multiple dimensions, and may be related to an F distribution by a final scaling [15]:

$$x_F = \frac{n_i + n_j - p - 1}{p(n_i + n_j - 2)} T^2 \sim F(p, n_i + n_j - p - 1) \tag{10.4}$$

It can be seen from this final relation that evaluation of the complement of the cumulative F-distribution function at x_F yields the p value for accepting the null hypothesis: the points in groups i and j are in fact drawn from the same distribution.

10.2.2 Dendrogram Generation

The implementation of the tree-generation procedure is a classical UPGMA algorithm [17]. When p values are reported at each branch point, a single tree is generated based on the matrix of Mahalanobis distances between groups. In the case of bootstrapped trees, the groups are randomly resampled with replacement while preserving group size. The desired number of trees is then generated using Euclidean distances between group means. The final tree used to report bootstrap

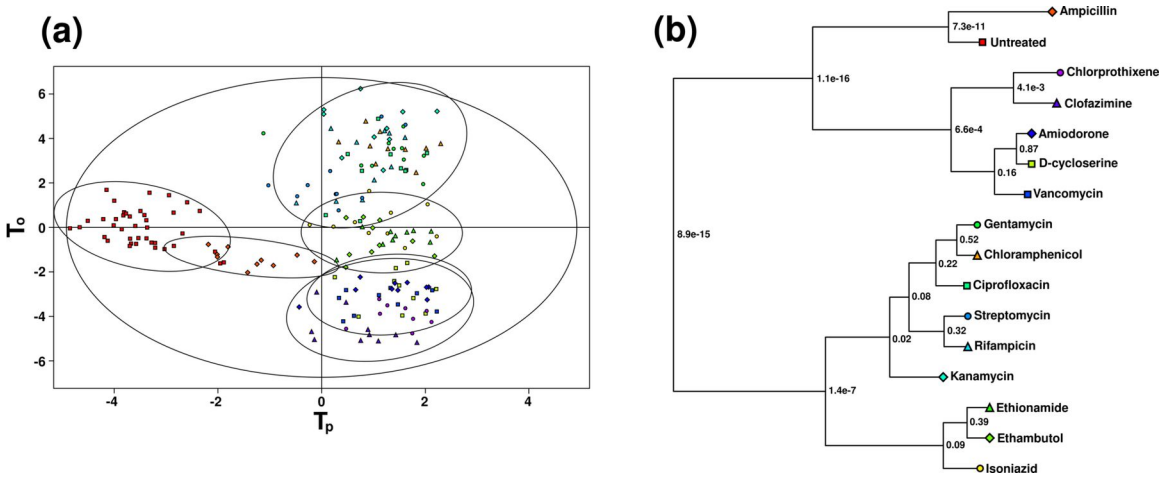

Figure 10.1: Confidence Ellipses and p-dendrogram of Example OPLS-DA Scores. (**A**) 2D OPLS-DA scores plot illustrating 95% confidence ellipses for a model having one predictive (PLS) and one orthogonal (OSC) component. The symbol shape and color each point correspond to the groups in (**B**). Discrimination in the first component is between wild-type and antibiotic-treated *Mycobacterium smegmatis*, and separations along the second component indicate metabolic differences between different antibiotic treatments. The antibiotics cluster together based on a shared biological target (cell wall synthesis, mycolic acid biosynthesis, or transcription, translation and DNA supercoiling). (**B**) Dendrogram generated from the scores in (**A**) using Mahalanobis distances, with p values for the null hypothesis reported at each branch.

probabilities is built using a Euclidean distance matrix calculated from the original (non-resampled) dataset.

10.2.3 Confidence Ellipse Calculation

When viewing PCA and PLS-DA scores plots, it was common practice to apply hand-drawn ellipses to inform group membership, or even to omit such ellipses entirely. This may lead to inconsistent or erroneous interpretation of experimental results. Instead, the fact that the Mahalanobis distances of a set of p-variate points from their sample mean follow a χ^2 distribution having p degrees of freedom [10] may be leveraged to estimate 95% confidence ellipses for scores in any number of dimensions. The sample mean $\mathbf{u}$ and sample covariance matrix $\mathbf{S}$ for each group must first be calculated from its scores-space data. Then, each group covariance matrix is decomposed into its eigenvalues and eigenvectors,

$$\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} \tag{10.5}$$

where $\mathbf{Q} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix holding the eigenvectors of $\mathbf{S}$, and $\mathbf{\Lambda} \in \mathbb{R}^{p \times p}$ is a diagonal matrix holding the corresponding eigenvalues of $\mathbf{S}$.

For the case of two-dimensional scores data, the 95% confidence ellipse for a group is as follows:

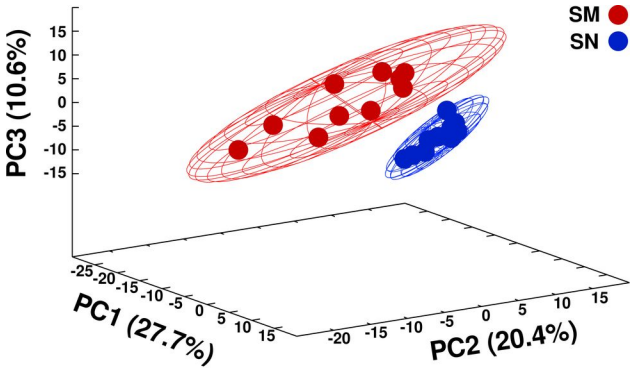

Figure 10.2: Confidence Ellipsoids from PCA Scores.
3D PCA scores plot with superimposed 95% confidence ellipsoids drawn as meshes containing group points. The ellipsoids define the statistical significance of class separation and provide an illustration where two groups are distinct in three-dimensional scores space.

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \mathbf{u} + \mathbf{Q}\sqrt{\mathbf{\Lambda} F^{-1}_{0.95,2}} \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} \tag{10.6}$$

where $F^{-1}_{0.95,2}$ is the value of the inverse χ^2 cumulative distribution function at $\alpha = 0.05$ and two degrees of freedom, and the square-root is taken element-wise over $\mathbf{\Lambda}$. Similarly, a three-dimensional (3D) confidence ellipsoid may be obtained from the following parametric equation:

$$\begin{bmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{bmatrix} = \mathbf{u} + \mathbf{Q}\sqrt{\mathbf{\Lambda} F^{-1}_{0.95,3}} \begin{bmatrix} \cos(u)\cos(v) \\ \cos(u)\sin(v) \\ \sin(v) \end{bmatrix} \tag{10.7}$$

where the parameters t, u and v are all evaluated on $(0, 2\pi)$. These methods allow for the inclusion of confidence regions onto two- and three-dimensional scores plots that reflect the 95% membership boundaries for each group. The approach assumes normally distributed within-group errors. Figures 10.1A and 10.2 illustrate the inclusion of these group confidence regions in representative PCA and OPLS-DA scores, respectively. The ellipses and ellipsoids clearly define statistically significant class separation and also provide an example in which multiple groups actually belong to the same underlying biological classification.

10.3 Results and Discussion

The described PCA utilities software package consists of a set of standalone C programs that generate dendrograms from PCA, PLS-DA and OPLS-DA scores, report p values and bootstrap numbers on tree branches, and incorporate confidence ellipses/ellipsoids into scores plots. The p values reported for every pair of distinct groups in scores space provide a truly quantitative means to discuss group separations. Support for the generation of dendrograms with these p values at each branch point is also included as an alternative answer to the bootstrap for answering the question of tree uniqueness. This eliminates the prior dependence on PHYLIP [16] reported for the original PCA-toTree [18] software package. The reporting of p values is complementary to bootstrapping methods

in cases of highly overlapped groups, in that it provides a more direct, interpretable quantitation of group separation.

In comparison with PCAtoTree, the PCA utilities software package now uses Mahalanobis distances because this metric is more appropriate for multivariate data. De Maesschalck et al. [3] provide an exceptional introduction to the use of Mahalanobis distances with PCA. Specifically, Mahalanobis distances account for different variances in each scores-space direction (t_1, t_2, t_3, etc.) and are invariant to scaling transformations. This accounting for variances-covariance structure ensures that the use of a Mahalanobis distance metric for dendrogram generation includes cluster shape and orientation in the analysis of group separation. Also, Mahalanobis distances calculated between groups in PCA scores space will closely approximate those calculated from the original data matrix while avoiding possible multicollinearities among the original variables. This is not true of Mahalanobis distances in PLS or OPLS scores space, because of the underlying supervision of the PLS algorithm. These features differ from the Euclidean distance metric, which is a special case of the Mahalanobis metric that arises when the group covariance matrices equal the identity matrix. Figure 10.1B illustrates the dendrogram structure based on the use of Mahalanobis distances determined from a set of scores, and Figure 10.3 shows the dendrogram structure based on Euclidean distances from the same scores.

It is important to note that our software is not a means of determining the reliability of PCA or PLS-DA models, but only a tool set for quantifying the scores that those models produce. In the case of PCA scores, significance of the principal components used must be inferred based on the explained sum of squares or another cross-validation technique [5, 13]. PLS-DA models require rigorous cross-validation to ensure model reliability, as they almost always yield perfect separations between the scores of different groups [11]. With that in mind, separations between groups not under discrimination may be due to true experimental differences in PLS-DA scores plots, as opposed to the forced separations between discriminated groups. Thus, the interpretation of any results from the PCA utilities must be done with the knowledge of the underlying algorithm's mathematical intent, and only after the model has been validated. While we demonstrated confidence region generation using only 2D and 3D scores plots, it is important to note that the PCA utilities software package places no restriction on the number of components or on which components may be used during dendrogram generation and p value calculation. Any dimensionality or choice of scores may be used with the described methods, provided all components are suitably validated.

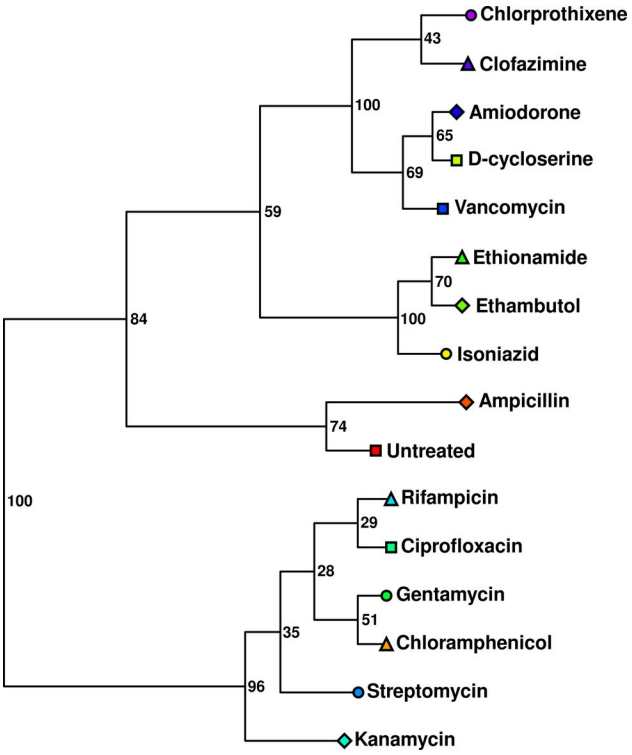

Figure 10.3: Dendrogram Generated using Euclidean Distances. Bootstrapped dendrogram generated from the scores data in Figure 10.1A using a Euclidean distance metric. Bootstrap statistics reported at each branch were computed from 5,000 bootstrap iterations.

The updated and enhanced version of PCAtoTree, called PCA utilities, provides a novel means of quantifying and visualizing separation significance in PCA, PLS-DA and OPLS-DA scores plots. Importantly, PCA utilities enables single-step methodologies for generating informative scores plots and dendrograms of experimental groups in *any* study utilizing PCA, PLS-DA or OPLS-DA to elucidate group structure in complex datasets, including metabolic fingerprinting and untargeted metabolic profiling. The tools are distributed under version 3.0 of the GNU General Public License [7] and are freely available at `http://bionmr.unl.edu/pca-utils.php`.

10.4 References

[1] P. E. Anderson, N. V. Reo, N. J. DelRaso, T. E. Doom, and M. L. Raymer. Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics*, 4(3):261–272, 2008.

[2] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227, 1979.

[3] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[4] S. J. Dixon, N. Heinrich, M. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevejo, and R. G. Brereton. Use of cluster separation indices and the influence of outliers: application of two new

separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles. *Journal of Chemometrics*, 23(1-2):19–31, 2009.

[5] H. T. Eastment and W. J. Krzanowski. Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*, 24(1):73–77, 1982.

[6] J. Felsenstein. PHYLIP – Phylogeny Inference Package (version 3.2). *Cladistics*, 5:164–166, 1989.

[7] F. S. Foundation. GNU General Public License, version 3. `http://www.gnu.org/licenses/gpl.html`, June 2007. Last retrieved 2015-05-11.

[8] A. M. Goodpaster and M. A. Kennedy. Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, 109(2):162–170, 2011.

[9] A. M. Goodpaster, L. E. Romick-Rosendale, and M. a. Kennedy. Statistical significance analysis of nuclear magnetic resonance-based metabonomics data. *Analytical Biochemistry*, 401(1):134–143, 2010.

[10] H. Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2:360–378, 1931.

[11] K. Kjeldahl and R. Bro. Some common misunderstandings in chemometrics. *Journal of Chemometrics*, 24(7-8):558–564, 2010.

[12] K. Koutroumbas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2006.

[13] W. J. Krzanowski. Cross-Validation in Principal Component Analysis. *Biometrics*, 43(3):575–584, 1987.

[14] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):7, 1936.

[15] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

[16] J. D. Retief. Phylogenetic analysis using PHYLIP. *Methods in Molecular Biology*, 132(2):243–258, 2000.

[17] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438, 1958.

[18] M. T. Werth, S. Halouska, M. D. Shortridge, B. Zhang, and R. Powers. Analysis of metabolomic PCA data using tree diagrams. *Analytical Biochemistry*, 399(1):58–63, 2010.

Chapter 11
Analysis of Protein $n - \pi^*$ Interactions

Whether you can observe a thing or not depends on the theory which you use. It is the theory which decides what can be observed.

– Albert Einstein

11.1 Introduction

Proteins exhibit a diversity of structures, with 2,738 folds or topologies present in the CATH database [14]. Each unique structure is defined by its amino acid composition, where sequence identities greater than 40% imply homologous structures [23]. Predicting the three-dimensional conformation of a protein from its primary sequence is a fundamental challenge of structural biology, and achieving this goal requires a thorough understanding of the underlying interactions and forces that stabilize protein structures [32].

Hydrophobic interactions and hydrogen bonds are two of the most common forces attributed to the overall stability of protein structures [22, 10]. The burial of hydrophobic residues is generally considered a major driving force in protein folding [13] and has been predicted to contribute roughly 8 kJ/mol per buried residue. Conversely, the contribution of hydrogen bonds to protein structure stability has been controversial [19]. Hydrogen bonds have been described as destabilizing, partially stabilizing or important driving forces. Of course, hydrogen bonds are a defining feature of α-helices, β-sheets and turns. Thus, the generally accepted view is that hydrogen bonds within a protein structure are marginally favored over hydrogen bonds to water. Hydrogen bonds are estimated to contribute roughly 4 kJ/mol to protein stability, but can vary based on the polarity of the microenvironment [12]. Despite these observations, a satisfying general mechanism for protein folding has not yet been described [25, 26], which strongly implies that our understanding of the factors involved in protein folding and stability is incomplete.

In a recent paper, Bartlett et al. proposed a new and potentially important interaction analogous to the hydrogen bond [3]. Unfortunately, the predicted $n - \pi^*$ interaction was based on density

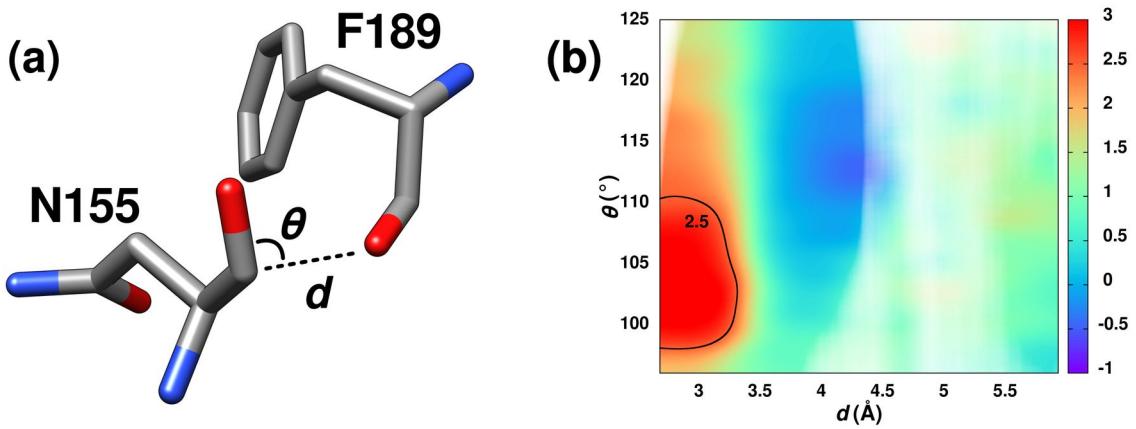

Figure 11.1: Predicted $n-\pi^*$ Interaction and Associated Carbonyl ^{13}C Chemical Shifts. (**A**) Residues Asn155 and Phe189 from the x-ray structure of *Bacillus amyloliquefaciens* subtillisin BPN' (PDB ID: 1v5i) illustrating the structural features for an optimal $n-\pi^*$ interaction between carbonyl groups. (**B**) 2D contour plot of carbonyl ^{13}C chemical shift differences relative to random coil values as a function of the distance (d) and angle (θ) between carbonyls. A Gaussian smoothing function was applied to the data with Δx and Δy of 0.3 Å and 1.5°, respectively. A transparency mask based on the density of experimental data (Figure 11.2) is overlaid on the contour plot. Regions lacking experimental data are white. Positive values indicate downfield shifts.

functional theory and a relatively low-level basis set. Conventional Kohn-Sham density functional theory does not properly model virtual orbitals [17] such as the π^* orbital proposed by Bartlett et al. to have a role in protein stabilization. Moreover, the relatively low-level basis set used by the authors is inadequate to model such orbitals, and likely gives rise to substantial basis-set superposition errors. Experimental data in support of this prediction was also not presented. Nevertheless, the predicted $n-\pi^*$ interaction was suggested to aid in the stabilization of protein structures and contribute roughly 0.4 to 5.4 kJ/mol. This stabilization was predicted to occur through the electron delocalization of the lone pair (n) of a carbonyl oxygen atom to the antibonding π^* orbital of a neighboring carbonyl carbon atom. An optimal $n-\pi^*$ interaction was predicted to be restricted to a specific range of structural parameters (Figure 11.1A) corresponding roughly to the Bürgi-Dunitz trajectory [5]. The distance (d) between the donor oxygen and acceptor carbon must be ≤ 3.2 Å, and the angle between the (donor O)$\cdots$(acceptor C) vector and the acceptor carbonyl vector, θ, must lie between 99° and 119°. Interestingly, the structural parameters required for an optimal $n-\pi^*$ interaction are prevalent in a wide variety of common secondary structures, including α-helices, 3_{10}-helices and twisted β-sheets, suggesting a potential alternative explanation.

Despite the presence of numerous conformations consistent with the $n-\pi^*$ interaction in protein structures, no experimental evidence was presented that supported the actual existence of this inter-

action. NMR chemical shifts of sp^2-hybridized groups contain a paramagnetic component caused by mixing of excited states with non-zero orbital angular momentum into the diamagnetic ground state [21]. These excitations are predominantly $n-\pi^*$ and $\pi-\pi^*$ and are therefore highly sensitive to perturbations of these orbitals. The predicted $n-\pi^*$ interactions between neighboring carbonyls would be expected to modify the local electronic environment of the acceptor carbonyl carbon nucleus, and the NMR ^{13}C chemical shift of the acceptor carbonyl carbon would experience a significant chemical shift change in the presence of an $n-\pi^*$ interaction [1]. Indeed, a strong (roughly 20 ppm range) linear relationship has been previously observed between carbonyl ^{13}C chemical shifts and the carbonyl $n-\pi^*$ transition energy [24, 8].

An extensive analysis of ^{13}C chemical shifts correlated to high-resolution x-ray structures combined with a detailed analysis of the molecular orbitals of a formamide trimer model complex does not support the postulated $n-\pi^*$ interaction. In fact, our model indicates that an $n-\pi^*$ interaction is implausible. Instead, a simple dipole-dipole interaction better explains the observed effects used in support of the $n-\pi^*$ interaction. While a prior manuscript by the same authors dismissed the dipole-dipole interaction explanation without elaboration [6], this work suggests it is a more plausible explanation of the observed data.

11.2 Materials and Methods

11.2.1 Analysis of Experimental Structures

A detailed statistical analysis was performed to correlate experimentally observed carbonyl ^{13}C chemical shifts with structural parameters between all possible pairs of carbonyls. Specifically, the angle between the carbonyls (θ) and the distance (d) between the oxygen and carbon were compared to experimental carbonyl ^{13}C chemical shifts. The PISCES [28] (`http://dunbrack.fcc.edu/pisces`) set of 2,885 high-resolution ($<$ 1.6 Å) x-ray crystal structures with less than 30% pairwise sequence identity selected from the RCSB Protein Data Bank (PDB) [4] was used for this analysis. Each structure was associated with assigned NMR ^{13}C and ^{15}N chemical shifts from the Biological Magnetic Resonance Bank (BMRB: `http://www.bmrb.wisc.edu`) [27] by FASTA [20] sequence alignments. The best match with an E-value $\leq 1.0 \times 10^{-9}$ and sequence identity $\geq 95\%$ was chosen, where the median E-value was 3.8×10^{-40} . The aligned ^{13}C and ^{15}N chemical shifts were uniformly referenced with the SHIFTCOR software tool [30]. The protein interfaces, surfaces and assemblies software tool (PISA, `http://www.ebi.ac.uk/pdbe/prot_int/pistart.html`) [15]

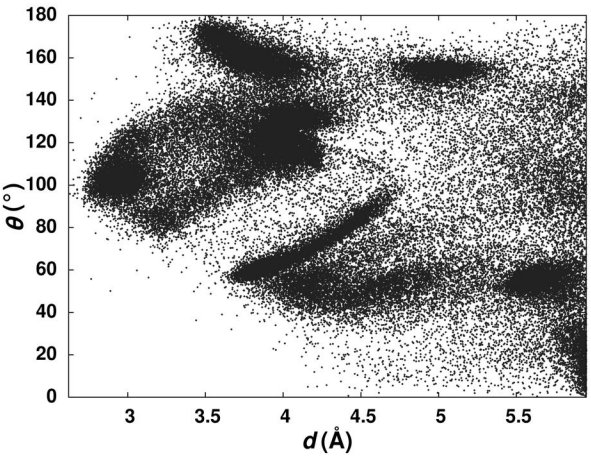

Figure 11.2: Population of (d, θ)-space by Experimental Structures. Plot of the distance (d) and angle (θ) measured between each of the 45,792 pairs of carbonyls with a potential $n - \pi^*$ interaction. The relative density of points in the occupied d and θ space was used to generate a transparency mask for Figure 11.1.

was used to predict residues involved in crystal packing interfaces. Residues with B-factors two standard deviations from the mean within each structure were identified as dynamic. Also, 3,699 NMR solution structures with PDB depositions cross-linked to the BMRB were used as a validation dataset, with alignments performed in an identical fashion to the analyzed x-ray structures.

A set of Perl scripts was written to extract structural parameters from the x-ray structures. For each structure in the selected set, all pairs of residues were analyzed for the possibility of an $n - \pi^*$ interaction. Values of d and θ were calculated for each residue pair, and torsional angles ϕ and ψ were calculated for the "acceptor" residue of each pair. Pairs of carbonyls with d and θ values within the optimal limits for an $n-\pi^*$ interaction were labeled as interactors (Figure 11.2). Standard random-coil chemical shifts were subtracted from the experimental carbonyl ^{13}C chemical shifts for each residue.

For all pairs of residues, a dipole-dipole potential (V_{dd}) was calculated from the high-resolution x-ray structures using equation 11.1:

$$V_{dd} = \frac{\vec{\mu}_1 \cdot \vec{\mu}_2 - 3(\vec{\mu}_1 \cdot \hat{r})(\vec{\mu}_2 \cdot \hat{r})}{4\pi\epsilon_0 |\vec{r}|^3} \tag{11.1}$$

where $\vec{\mu}_1$ and $\vec{\mu}_2$ are the two C=O bond vectors, $\vec{r}$ is the vector between the centers of the C=O bonds, and $\hat{r}$ is its unit vector. The nominal value of 2.34 Debye was taken for the carbonyl dipole moment. Similarly, for all residue pairs, the minimum possible hydrogen bond length (d_{O-H}) was calculated from the high-resolution x-ray structures. Hydrogen bond lengths were calculated based on the nearest non-neighboring backbone amide hydrogen, with a maximal bonding angle of 60°. Figure 11.3 illustrates the relationship between V_{dd} and ^{13}C chemical shifts of all carbonyl pairs in

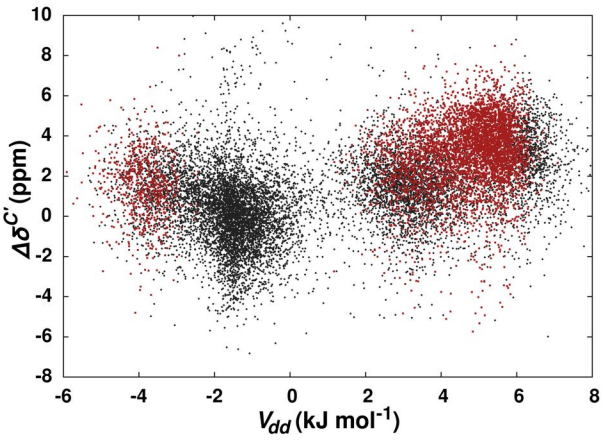

Figure 11.3: Carbonyl ^{13}C Chemical Shifts and Dipole-Dipole Potential. Carbonyl ^{13}C chemical shift differences relative to random coil are plotted against calculated dipole-dipole potential (V_{dd}). The dipole-dipole potential is calculated from the high-resolution x-ray structure using equation 11.1. Pairs of carbonyls with d and θ values within the optimal limits for an $n-\pi^*$ interaction are colored red.

the dataset.

11.2.2 Model Compound Calculations

Quantum chemical calculations were performed using the program Gaussian-09 [11]. A nearly planar formamide head-to-tail dimer, composed of a formamide monomer (molecule 1) hydrogen bonded through its C=O group to the N-H group of a second, nearly parallel formamide (molecule 2) was chosen to approximate the hydrogen bonding motif found in both α-helices and antiparallel β-sheets. The dimer was fully optimized at the MP2/6-311++G(2d,p) level; Möller-Plesset second order perturbation theory (MP2) was chosen because it is superior in modeling long-range and dispersive contributions to the electron correlation Hamiltonian. A third formamide (molecule 3) was then added to generate the putative $n-\pi^*$ interaction with molecule 1, imposing the following constraints: (1) $C_3{=}O_3\cdots C_1$ angle fixed at 90°, to ensure the n_π orbital of molecule 3 points toward the carbonyl of molecule 1 (2) $O_3\cdots C_1{=}O_1$ constrained to a set of fixed angles θ, ranging from 70° to 120° (3) $O_3\cdots C_1$ constrained to a set of fixed distances d, ranging from 2.9 Å to 3.3 Å (5) $O_1\cdots N_2$ constrained to a set of fixed distances, ranging from 2.8 Å to 3.2 Å, to vary the strength of the hydrogen bond. The system of three molecules was then subjected to constrained optimization at the same level as before. The optimized trimolecular system at an angle $\theta = 90°$ is shown in Figure 11.4. Finally, a full set of shielding tensors was computed using standard Gauge-independent atomic orbital methods.

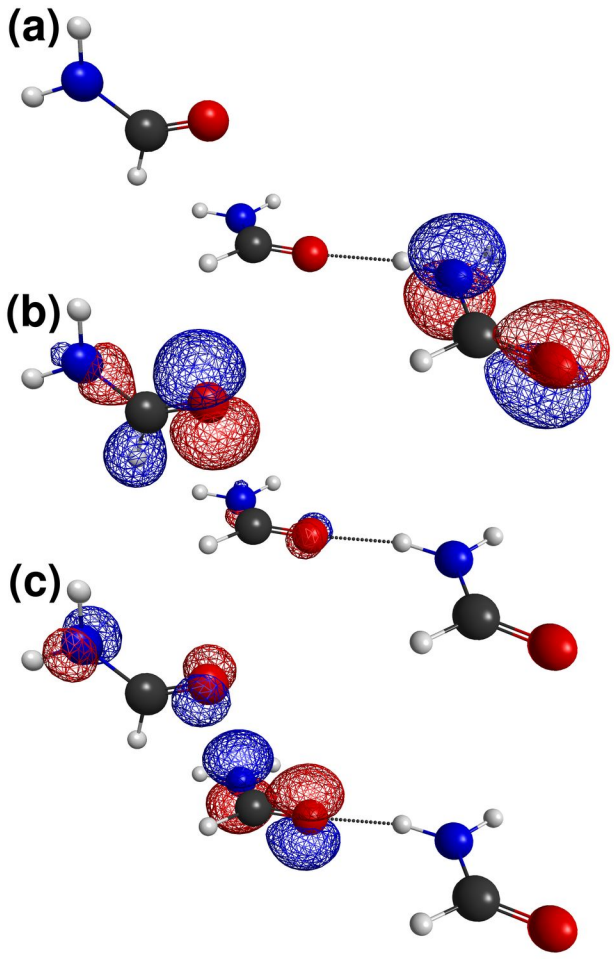

Figure 11.4: Formamide Trimer Model.
Molecular orbitals of (**A**) the hydrogen bond donor, (**B**) the putative $n_\pi - \pi^*$ donor and (**C**) the putative $n_\pi - \pi^*$ acceptor, in the trimeric complex used in quantum chemical calculations.

11.3 Results

A total of 2,516,360 residue pairs from a set of 164 high-resolution (< 1.6 Å) x-ray crystal structures with assigned and uniformly referenced carbonyl ^{13}C chemical shifts were analyzed for potential $n - \pi^*$ interactions. Setting a maximal distance of 6.0 Å between the donor oxygen and acceptor carbon yielded 45,792 potential acceptor carbonyl carbon atoms. The carbonyl ^{13}C chemical shift differences relative to random coil values for each of the 45,792 potential acceptor carbonyls were plotted against the d and θ values for each carbonyl pair (Figure 11.1B). These chemical shift differences represent the contribution from the local structural environment, and potentially the $n - \pi^*$ interaction. The two-dimensional contour plot indicates a maximal downfield shift of 2.9 ppm centered on the optimal structural parameters predicted for an $n - \pi^*$ interaction.

Of the 45,792 carbonyls, 5,378 had optimal d and θ values for an $n - \pi^*$ interaction and 40,414 were outside this optimal range. The mean carbonyl ^{13}C chemical shift difference for the 40,414 carbonyls labeled as non-interactors is 0.58 ± 1.98 ppm. In contrast, the mean carbonyl ^{13}C chemical shift difference for the 5,378 interactors is 2.93 ± 2.41 ppm. A Student's t-test indicates the difference of 2.35 ppm between the two means is statistically significant at the 99.9% confidence level. To address possible errors introduced into the analysis by highly dynamic residues in the x-ray structures, possible carbonyl interactors with B-factors greater than two standard deviations above the mean were omitted from the dataset. In the resultant set of 44,302 potential carbonyl interactors, the 2.33 ppm chemical shift difference was statistically indistinguishable from the original analysis. Similarly, possible interactors predicted at a 95% confidence level to participate in crystal-packing interfaces were also omitted from the dataset. Again, the corresponding set of interactors yielded a chemical shift difference of 2.33 ppm, which is still statistically significant at the 99.9% confidence level.

To address the possibility that differences between x-ray crystal structures and NMR solution structures may lead to errors in the analysis, a replicate analysis was performed on a set of 137 NMR solution structures corresponding to the same set of ^{13}C and ^{15}N chemical shifts used previously. Structural alignments using MAMMOTH showed a mean RMSD of 1.87 ± 0.57 Å between the pairs of x-ray and NMR structures. Of the 1,419,547 resulting carbonyl pairs from the NMR structures, 38,534 pairs were found to be potential interactors. Of the carbonyls in that set, 2,510 interactors were found with a mean carbonyl ^{13}C chemical shift difference of 2.84 ± 1.71 ppm. The remaining

36,024 non-interactors had a mean carbonyl ^{13}C chemical shift difference of 1.02 ± 2.02 ppm. Again, the 1.82 ppm difference between the two means is statistically significant at the 99.9% confidence level, indicating that differences between x-ray and NMR structures cannot account for the observed downfield ^{13}C chemical shift.

As predicted, a clear correlation is observed between structural regions consistent with an optimal $n-\pi^*$ interaction and a downfield shift of the accepting carbonyl ^{13}C resonance. Interestingly, the potential $n-\pi^*$ interactions were primarily observed between sequential ($|i-j|=1$) carbonyls. Out of the 164 structures and 2,516,360 residue pairs, only four pairs of carbonyls exhibited a through-space ($|i-j|>5$) arrangement consistent with an optimal $n-\pi^*$ interaction. This result implies any protein stabilization energy obtained from the proposed $n-\pi^*$ interaction is opportunistic, as opposed to a driving force for protein folding. Apparently, the formation of through-space $n-\pi^*$ interactions is simply less favorable than for other interactions, such as hydrogen bonds or salt-bridges. This also implies that the predicted energy of 5.4 kJ/mol for an optimal $n-\pi^*$ interaction is an over-estimate.

In actuality, an $n-\pi^*$ interaction that imparts a stability of 5.4 kJ/mol would likely fix adjacent pairs of carbonyl groups to preferred torsional angles in order to maximize this interaction. Correspondingly, the existence of these highly energetic $n-\pi^*$ interactions would likely be detrimental to properly folding a protein structure. Folding a protein to its native fold would require distorting the majority of carbonyl pairs away from the ideal torsion angles for a proper $n-\pi^*$ interaction. Only 12% (5,378 out of 45,792) of carbonyls from the 164 x-ray structures adopted conformations with optimal d and θ values for an $n-\pi^*$ interaction. As a result, folding every protein structure would incur an initial energetic penalty of nearly 5.4 kJ/mol per carbonyl pair.

A predominant number of the carbonyls consistent with an optimal $n-\pi^*$ interaction and with a downfield shift of roughly 2.5 ppm fall within the typical α-helical region of a Ramachandran plot, where the remaining residues are near the twisted β-sheet region (Figure 11.5). Significant chemical shift changes for carbonyl residues within secondary structures are well documented [29]. Previous analyses of structural factors contributing to carbonyl ^{13}C chemical shifts have implicated hydrogen bond formation [9, 2, 31] or excluded hydrogen bond formation [7, 18, 16], have implicated ϕ, ψ, and χ dihedral angles [18] or have excluded secondary structure parameters [7, 9]. Thus, other factors, such as hydrogen bonds or dipole-dipole interactions, may explain the apparent correlation between carbonyl ^{13}C shifts and the optimal d and θ values for an $n-\pi^*$ interaction. This is probable

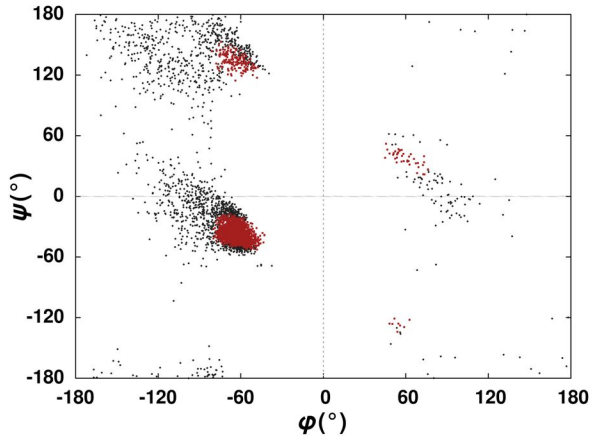

Figure 11.5: Population of (ϕ, ψ)-space by Experimental Structures. Ramachandran plot of carbonyls with ^{13}C chemical shift differences relative to random coil that are > 2.5 ppm. The acceptor carbonyls from each pair of carbonyls with d and θ values within the optimal limits for an $n - \pi^*$ interaction are colored red.

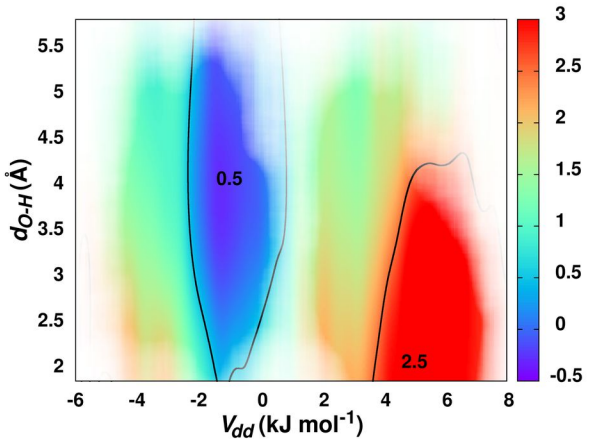

Figure 11.6: Carbonyl ^{13}C Chemical Shifts and Hydrogen Bonds. Contour plot of ^{13}C carbonyl chemical shift differences as a function of calculated dipole-dipole potential (V_{dd}) and calculated hydrogen bond length (d_{O-H}).

given the association of $n - \pi^*$ interactions with secondary structure elements. The contribution of a dipole-dipole interaction to carbonyl ^{13}C chemical shifts is illustrated in Figure 11.3. The dipole-dipole potentials were calculated using the high-resolution x-ray structures for each of the 45,792 carbonyl pairs with a maximal distance of 6.0 Å between the donor oxygen and acceptor carbon. While there is significant scatter in the data, there is also a clear trend between a downfield carbonyl ^{13}C chemical shift and an increasing dipole-dipole energy. Importantly, the cluster of acceptor carbonyls in Figure 11.3 with the largest ^{13}C chemical shift difference (3.15 ± 2.44 ppm) and positive dipole-dipole potentials also conforms to the optimal d and θ values for the predicted $n - \pi^*$ interaction.

The contribution of a hydrogen-bond interaction to the carbonyl ^{13}C chemical shift was similarly evaluated by calculating the shortest oxygen-hydrogen distance (d_{O-H}) for each donor carbonyl. Again, the distances were calculated using the high-resolution x-ray structures for each of the 45,792 carbonyl pairs. A three-dimensional plot comparing the dipole-dipole potentials, oxygen-hydrogen

distances, and the associated carbonyl ^{13}C chemical shifts is very revealing. It can be clearly seen from Figure 11.6 that any contribution from a hydrogen bond to the ^{13}C carbonyl chemical shift is minimal relative to the dipole-dipole contribution. Both the α-helical and β-sheet regions, which obviously contain hydrogen bond interactions, have distinctly different ^{13}C carbonyl chemical shifts. The α-helical region corresponds to a positive dipole-dipole interaction, and correspondingly to a large carbonyl ^{13}C chemical shift difference. Conversely, the β-sheet region has a negative dipole-dipole interaction and a near zero carbonyl ^{13}C chemical shift difference. These results further indicate a consistency with a dipole-dipole interaction as opposed to the predicted $n-\pi^*$ interaction.

It is important to note that there is a second cluster of carbonyls in Figure 11.3 with low ^{13}C chemical shifts and negative dipole-dipole potentials that are also consistent with the optimal d and θ values for the predicted $n-\pi^*$ interaction. A visual inspection of the x-ray structures indicates that these carbonyl pairs are actually pointing away from each other and do not form the configuration for an $n-\pi^*$ interaction illustrated in Figure 11.1A. Clearly, d and θ values alone fail to adequately define the optimal geometry of the dipole-dipole interaction that is apparently responsible for the observed downfield ^{13}C chemical shifts.

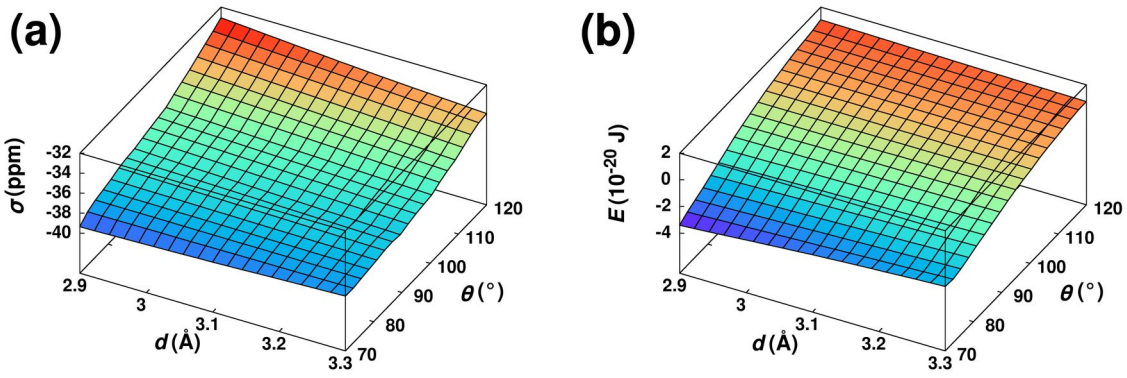

Figure 11.7: Summary of Quantum Chemical Calculations.
Plot of calculated (**A**) carbonyl ^{13}C chemical shielding (σ) and (**B**) dipole-dipole interaction energy (E) as a function of the distance between donor oxygen and acceptor carbon (d) and the angle between carbonyl groups (θ).

To further examine the origin of these effects, quantum chemical calculations were conducted on a model system, a formamide trimer in which molecules 2 and 3 form an approximately planar, head-to-tail hydrogen bonded dimer, and molecule 3 acts as a putative $n-\pi^*$ donor, with the n_π "donor" oxygen fixed at a distance d which ranges from 2.9 Å and 3.3 Å from the carbonyl carbon of

molecule 2, with the $O_3 \cdots C_2$ vector also fixed at angles θ from 70° to 120° from the $C_2{=}O_2$ vector. To avoid problems with the use of density functional theory to model virtual orbitals, Möller-Plesset second order perturbation theory (MP2) was instead used, with a substantially larger basis set than in the previous work. The geometry and relevant Hartree-Fock orbitals of the complex used is shown in Figure 11.4, for $d = 2.9$ Å and $\theta = 100°$. The computed chemical shielding is shown in Figure 11.7A as a function of d and θ. The shielding decreases monotonically with θ, but, in contrast, the slope of the shielding surface with respect to d changes sign between $\theta = 70°$ and $\theta = 120°$. This shielding surface does not have the geometry expected if the chemical shielding dependencies on θ and d were dominated by an $n_\pi - \pi^*$ interaction, where shielding should be maximal at θ slightly larger than 90° and $d = 2.9$ Å, decreasing rapidly with increasing values of d.

However, the shielding surface does show a remarkable similarity to the dipole-dipole energy between the putative donor and acceptor, as shown in Figure 11.7B. This energy was computed using a very simple model assuming the electric dipole vector lies along the carbonyl bond for both molecules and has a value of 2.34 D or 7.81×10^{-30} C·m. As can be seen, the dipole energy closely parallels the chemical shielding surface, monotonically increasing with θ and inverting its slope with respect to d as θ increases. This indicates the major influence on the carbonyl ^{13}C chemical shielding is not an $n_\pi - \pi^*$ interaction but rather the electrostatic field from the neighboring carbonyl dipole. The correspondence is not, however, exact: the chemical shielding surface shows a small negative inflection around $\theta = 90°$, which is actually slightly reversed in the dipolar energy plot.

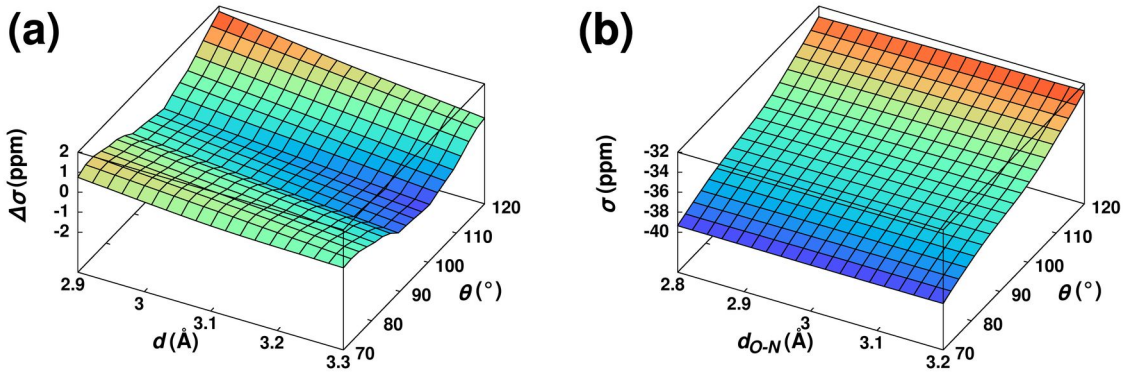

Figure 11.8: Supplemental Quantum Chemical Results.
(**A**) Plot of the residuals for the fit of the chemical shielding surface to a function proportional to the dipole-dipole energy. (**B**) Summary of the quantum chemical calculations of the hydrogen bond contribution to the dipole-dipole interaction; plot of carbonyl ^{13}C chemical shielding (σ) as a function of the hydrogen bond angle (θ) and distance (d_{O-N}).

In order to examine whether an $n_\pi - \pi^*$ interaction might be responsible for this inflection, the chemical shielding surface was fit to a function proportional to the dipolar energy, under the assumption the dipole moment vector lies along the C=O bond direction, and the best fit subtracted from the chemical shielding surface (Figure 11.8A). The residual shows a minimum at $\theta \sim 95°$, as would be expected for an $n_\pi - \pi^*$ interaction, but the dip does not appear to decrease rapidly as d increases, as an orbital overlap term would. In fact, the residual is slightly larger at $d = 3.3$ Å than at 2.9 Å (1.3 ppm vs. 1.1 ppm).

From the fit of the shielding surface to the estimated dipole interaction energy, with the assumption the magnitude of the electric dipole moment is that of a formamide monomer (3.7 D), a dependence of chemical shielding on field of −190 ppm/a.u. was obtained (1 atomic unit (a.u.) of electric field equals 5.142×10^{11} V/m). Direct calculations of the dependence of the shielding of an isolated formamide on an external applied field along the C=O bond direction gave a value of −150 ppm/a.u. However, it is highly likely that this estimation of the dipole-dipole interaction for two amides is low. Firstly, higher electric multipole terms were neglected in the calculation, and these are likely to be substantial for a moiety as asymmetric as a peptide linkage, at these close proximities. Second, the interaction of the dipoles is likely to be enhanced by the highly polarizable hydrogen bond, which is necessarily omitted in the monomer model. Agreement of the model with direct estimates of the effect of electric field on shielding is therefore rather good.

The dependence of chemical shielding on hydrogen bonding strength for all combinations of d and θ was examined as a function of the hydrogen bond distance d_{O-N} (Figure 11.8B). In accordance with the results of Wishart and others [7, 18, 16], and contrary to initial naïve expectations, the effect was very small and independent of the position of the putative $n_\pi - \pi^*$ donor carbonyl.

For the sake of completeness, the effect of an "end-on" carbonyl-carbonyl interaction was examined, using a dimeric cluster in which the "donor" carbonyl bond was parallel to the "donor" oxygen-"acceptor" carbon vector, resulting in a possible $n_\sigma - \pi^*$ interaction. As can be seen in Figure 11.9, for values of d ranging from 2.9 Å to 4.1 Å, the chemical shielding also follows the negative of the dipolar interaction energy over the range $70° < \theta < 120°$, with little evidence of any effect of orbital overlap on chemical shielding.

One other outcome of the calculations is of possible note. While there was little discernible ef-

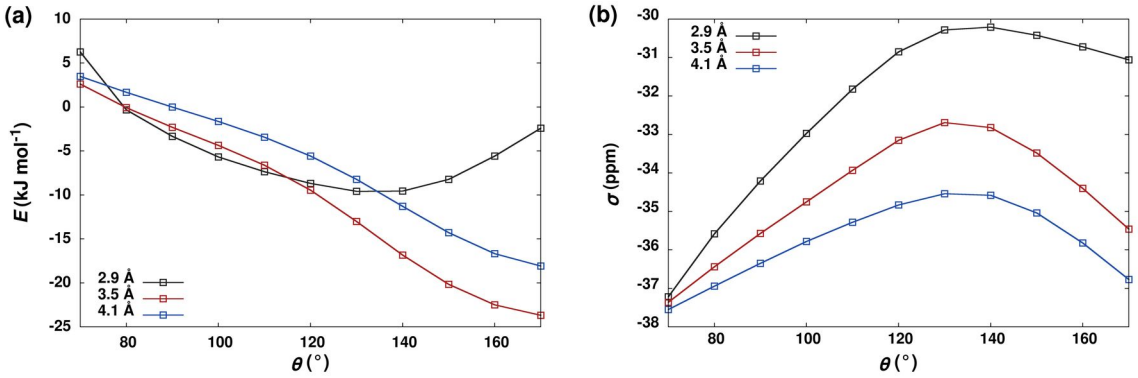

Figure 11.9: Summary of Quantum Chemical Calculations for "End-On" Dipole Interaction.
Plots of (**A**) interaction energy and (**B**) carbonyl ^{13}C chemical shielding (σ) as a function of the angle between the carbonyls (θ) for three different distances (d) between the donor oxygen and acceptor carbon.

fect of the proposed $n_\sigma - \pi^*$ or $n_\pi - \pi^*$ interactions on the shielding of the carbonyl carbon or the length of the carbonyl bond, substantial pyramidalization of the amide nitrogen was observed at low values of d and values of θ close to 90°. This would indicate that the primary effect of the "donor" carbonyl might not be on the carbonyl π bond per se, but on its delocalization over the entire amide group. There was also a substantial lengthening of the carbon-nitrogen bond – consistent with a reduced bond order – accompanied by substantial changes in the computed ^{15}N chemical shielding. Thus, while no evidence was found of effects from $n_\sigma - \pi^*$ interactions on the ^{13}C NMR spectroscopy or the energetics of the system, such interactions might be detectable in ^{15}N chemical shifts. Unfortunately, ^{15}N shifts are known to be much more dispersed than carbonyl ^{13}C shifts and are susceptible to a wide range of influences, so disentangling the interaction in real proteins might be a Herculean task.

11.4 Discussion and Conclusions

When the molecular orbitals for the trimeric complex are examined in detail, the above results become clear. It is in fact misleading to think of amide groups as being dominated by the carbonyl π bond. The highest occupied molecular orbital (HOMO) of the formamide trimer in fact consists almost entirely of p_z orbitals on the N and O, with wavefunctions of opposite sign. This is depicted in Figure 11.4A for the Hartree-Fock HOMO of the hydrogen bond donor (energy = −0.377 Ha). The orbital is slightly bonding with respect to the carbonyl, but the carbonyl carbon overall has very little contribution to the molecular orbital. The equivalent orbital of the putative acceptor

(Figure 11.4B) has somewhat lower energy (−0.438 Ha) but shows remarkably little mixing with other molecular orbitals, and in particular little mixing with the n_π orbital of the putative $n_\pi - \pi^*$ donor (Figure 11.4C). That orbital has in fact a very similar energy (−0.465 Ha), and at other geometries – specifically lower values of θ, mixes with the HOMO of the acceptor. The reason for this is quite simple: because the HOMO has only a very small contribution for carbonyl carbon orbitals, bringing the n_π orbital closer to it has very little effect. The mixing that is present at smaller values of θ in fact seems to be partly responsible for the increased pyramidalization of the nitrogen of the acceptor at those orientations. We see no evidence of any orbital mixing that could be attributed to $n_\pi - \pi^*$ interactions. Given the weakness of the mixing with orbitals that are very close in energy to n_π it is implausible that substantial mixing would be observed with an orbital almost a Hartree higher in energy.

In conclusion, quantum chemical calculations, experimental carbonyl ^{13}C chemical shifts and structural data indicate that a simple electrostatic dipole-dipole interaction explains the large downfield carbonyl ^{13}C chemical shift in an α-helix. There is no evidence for a significant contribution from an $n - \pi^*$ interaction to the carbonyl bond. The single indication of $n - \pi^*$ interactions seems to be a substantial lengthening of the carbon-nitrogen bond and pyramidalization of the nitrogen at θ angles favorable for these interactions. In fact, such pyramidalization seems to be a logical consequence of the electronic structure of amides, whose π orbitals are delocalized over the whole system.

11.5 References

[1] A. Abragam. *The Principles of Nuclear Magnetism*. Oxford University Press, 1961.

[2] N. Asakawa, S. Kuroki, H. Kurosu, I. Ando, A. Shoji, and T. Ozaki. Hydrogen-Bonding Effect on ^{13}C NMR Chemical-Shifts of L-Alanine Residue Carbonyl Carbons of Peptides in the Solid State. *Journal of the American Chemical Society*, 114(9):3261–3265, 1992.

[3] G. J. Bartlett, A. Choudhary, R. T. Raines, and D. N. Woolfson. $n-\pi^*$ Interactions in Proteins. *Nature Chemical Biology*, 6(8):615–620, 2010.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[5] H. B. Burgi, J. D. Dunitz, and E. Shefter. Geometrical Reaction Coordinates .2. Nucleophilic Addition to a Carbonyl Group. *Journal of the American Chemical Society*, 95(15):5065–5067, 1973.

[6] A. Choudhary, D. Gandla, G. R. Krow, and R. T. Raines. Nature of Amide Carbonyl-Carbonyl Interactions in Proteins. *Journal of the American Chemical Society*, 131(21):7244–7246, 2009.

[7] F. Cisnetti, K. Loth, P. Pelupessy, and G. Bodenhausen. Determination of Chemical Shift Anisotropy Tensors of Carbonyl Nuclei in Proteins through Cross-Correlated Relaxation in NMR. *ChemPhysChem*, 5(6):807–814, 2004.

[8] J. W. H. De. Relation between ^{13}C and ^{17}O carbonyl chemical shifts for $n - n^*$ transition energies. *Molecular Physics*, 18:31–37, 1970.

[9] A. C. de Dios, J. G. Pearson, and E. Oldfield. Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *Science*, 260(5113):1491–1496, 1993.

[10] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[11] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. M. Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian-09, Revision A.02, 2009.

[12] J. Gao, D. A. Bosco, E. T. Powers, and J. W. Kelly. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature Structural and Molecular Biology*, 16(7):684–690, 2009.

[13] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[14] M. Knudsen and C. Wiuf. The CATH database. *Human Genomics*, 4(3):207–212, 2010.

[15] E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica*, 60(12 I):2256–2268, 2004.

[16] P. R. L. Markwick and M. Sattler. Site-Specific Variations of Carbonyl Chemical Shift Anisotropies in Proteins. *Journal of the American Chemical Society*, 126:11424–11425, 2004.

[17] H. Mera and K. Stokbro. Using Kohn-Sham density functional theory to describe charged excitations in finite systems. *Physical Review*, 79:125109, 2009.

[18] S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart. Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *Journal of Biomolecular NMR*, 26:215–240, 2003.

[19] C. N. Pace. Energetics of protein hydrogen bonds. *Nature Structural and Molecular Biology*, 16(7):681–682, 2009.

[20] W. R. Pearson. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, 132:185–219, 2000.

[21] N. F. Ramsey. Magnetic Shielding of Nuclei in Molecules. *Physical Review*, 78(6):699–703, 1950.

[22] A. D. Robertson and K. P. Murphy. Protein Structure and the Energetics of Protein Stability. *Chemical Reviews*, 97(5):1251–1267, 1997.

[23] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

[24] A. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[25] E. Shakhnovich. Protein folding roller coaster, one molecule at a time. *PNAS*, 106(29):11823–11824, 2009.

[26] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330:341–346, 2010.

[27] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, W. R. Kent, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:402–408, 2008.

[28] G. Wang and R. L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[29] Y. Wang and O. Jardetzky. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Science*, 11(4):852–861, 2002.

[30] D. S. Wishart, H. Y. Zhang, and S. Neal. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25(3):173–195, 2003.

[31] B. J. Wylie, L. J. Sperling, H. L. Frericks, G. J. Shah, W. T. Franks, and C. M. Rienstra. Chemical-Shift Anisotropy Measurements of Amide and Carbonyl Resonances in a Microcrystalline Protein with Slow Magic-Angle Spinning NMR Spectroscopy. *Journal of the American Chemical Society*, 129:5318–5319, 2007.

[32] Y. Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.

Chapter 12
Summary and Future Directions

> *Chemists, in particular, cannot understand why they should fund someone to do data analysis.*
>
> – Richard Brereton [6]

12.1 The Need for Data Handling

The field of chemometrics is still in its infancy, but the chemometric practice of extracting quantitative chemical information from data collected on complex samples is much older, and has innumerable applications in chemistry [47, 6]. While the standard toolbox of t-tests, run charts and univariate distributions has served analytical chemists well, the analysis of spectral measurements of multicomponent mixtures demands a more computationally intensive approach.

However, optimal chemometric modeling of spectral data does not begin when the data are read in for the first time, but before acquisition has even been performed. Successful experimental design relies on data collection procedures that yield informative, high-quality measurement results. Spectra having the highest possible resolution, dynamic range and signal-to-noise ratio are necessary if reliable conclusions are to be drawn from their models. In multidimensional NMR experiments, methods of sparse data collection are becoming increasingly popular, as they provide avenues for maximizing spectral quality. In these nonuniform sampling (NUS) methods, the greatest contributing factor to spectral quality is the sampling scheme, and the generation of sampling schemes that optimize various spectral features (i.e. sensitivity or resolution) is still an active area of fundamental research [31]. Chapter 2 introduces a general framework for multidimensional nonuniform sampling that extends the work of Hyberts and Wagner [25] and deterministically generates nonuniform sampling schedules that perform as well or better than stochastic methods [53]. By suggesting an alternative mechanism for introducing irregularity into a sampling schedule, burst-augmented gap sampling aims to provoke further investigation into which features of a sampling schedule yield optimal spectral results. Furthermore, this new framework is the first proposed mechanism for deterministically

constructing sampling schedules on a multidimensional Nyquist grid [19] based on a general equation.

The processing, treatment and modeling of spectral measurements using multivariate statistics, outlined in Chapter 3, is a nuanced task, with many pitfalls awaiting the chemist who lacks experience and training in multivariate data analysis [49, 55]. Most applications of chemometrics are performed by analytical chemists, whose expertise lies with a certain type of instrumentation rather than statistics. In order to promote statistically sound data handling practices, chemometricians must begin to place easy-to-use, well-documented software packages in the hands of chemists. These software packages must simultaneously provide powerful mechanisms of multivariate data analysis, educate users about proper data handling, and encourage further extension and collaboration between fundamentally focused chemometricians and applications-driven chemists. Chapter 5 introduces MVAPACK [50], an open-source suite of simple GNU Octave [18] functions that aims to address those goals, and Chapter 4 describes its use on a wide variety of applications within the rapidly growing field of metabolomics [50, 30, 55]. The release of MVAPACK under an open-source license ensures transparency, allows for critical review by expert members of the chemometrics community, and enables modification and extension by its user base.

Without a doubt, the availability of MVAPACK made the development of phase-scatter correction (PSC, Chapter 6), uncomplicated statistical spectral remodeling (USSR, Chapter 7), and generalized adaptive intelligent binning (GAI-binning, Chapter 8) substantially easier [51, 56, 54]. By interlacing these methods into the existing fabric of MVAPACK, only a single new function had to be written for each new method. The data structures required by the algorithms – complex data matrices, real data matrices, and arrays of real matrices – are provided in a well-defined format by existing functions in MVAPACK. As a result, development could be focused 100% on functionalities of the *actual algorithms*, and not the "glue code" typically required to make a method even moderately useful. This modularity conveys distinct advantages to the entire community: from the perspective of an MVAPACK user, adding PSC or GAI-binning into an existing data handling protocol requires changing a single function call in an Octave script.[1]

As more chemists begin to collect multiple analytical measurements from each sample, it is imperative that well-defined, statistically acceptable methods be available to model the resulting data [45, 46, 38, 30, 52]. Consensus PCA (CPCA-W) and Multiblock PLS (MB-PLS), discussed in Chap-

[1] After upgrading to the latest version of MVAPACK, of course.

ter 3, are powerful extensions of PCA and PLS modeling to multiblock datasets, and are implemented in MVAPACK to provide easy access by the community. However, at the time of their implementation, no analogous extension of OPLS existed to handle multiblock data. Most chemists using multivariate statistics prefer OPLS over PLS due to its enhanced interpretability [42, 41], and the lack of a formally defined multiblock analog of OPLS was starting to become apparent in the form of several ad hoc attempts to use single-block OPLS on multiple matrices [10, 5]. Chapter 9 formally defines Multiblock OPLS (MB-OPLS) as a consensus bilinear factorization method, relates it to the OnPLS method proposed by Löfstedt and Trygg [29], and describes its relationship to several other methods (CPCA-W, MB-PLS, nPLS and OnPLS) in the context of the highly general framework described by Hanafi and Kiers [22]. Coupled with the inclusion of MB-OPLS in MVAPACK, this description presents a thoroughly vetted avenue for chemists to easily model their multiblock data using OPLS, without resorting to ad hoc approaches [52].

However, bilinear factorizations like PCA, PLS and OPLS merely represent *the very first step* towards chemometric modeling of spectral data [20]. Ideal chemometric models of spectral measurements would directly report the concentrations of the individual components in the mixtures being studied [17]. In the context of bilinear modeling, achieving this goal requires the imposition of stronger constraints on the model scores and loadings (cf. Section 3.5). Methods such as multivariate curve resolution by alternating least squares (MCR-ALS, [15]) and molecular factor analysis (MFR, [17]) impose non-negativity constraints on both the scores and loadings in an alternating least squares framework, while Bayesian spectral decomposition (BSD, [35, 40]) and Bayesian positive source separation (BPSS, [32, 33]) do so by assigning prior probabilities to their values. While these methods report more chemically and spectroscopically relevant information than PCA by imposing non-negativity, they still return mixtures of multiple compounds in their loadings. Imposition of "hard modeling" constraints takes the problem a step further by requiring each extracted signal to obey a certain parametric form. Hard modeling of NMR data has been accomplished using time-domain Bayesian [7, 8, 9, 12] and maximum-likelihood [13] modeling, as well as hybrid time- and frequency-domain maximum-likelihood [14, 11, 24] modeling. Such methods translate the task of identifying mixture components into one of peak-matching, where each signal in the model is assigned to a known set of signals from a given compound. Inclusion of compound identity information in the modeling process is the final step towards directly decomposing complex mixtures into their constituent parts. Methods such as BATMAN [1, 23] and BQuant [57] have been shown to perform quite capably in modeling 1D ^{1}H NMR spectra, but require the specification of spectral information

for each potential mixture component, and tend to be computationally expensive. Despite the clear advantage these more complex approaches hold over soft modeling, their adoption by the chemistry community has proceeded incredibly slowly. Without easy-to-use software implementations, these advanced modeling algorithms are likely to remain a mere afterthought in applied fields like metabolomics, where usability often outweighs capability.

Once models have been constructed around a dataset, the task of inference begins, usually involving a great deal more expert chemical or biochemical insight than model construction required. At this stage, it can be tempting to draw conclusions based on patterns observed in the results. The tendency of human perception to over-fit patterns to data can, however, lead analysts to infer too much from a model. When chemical conclusions must be drawn from scores plots of PCA, PLS-DA or OPLS-DA models, there are simple statistical measures that can be taken to avoid mis-perception of scores-space patterns. Chapter 10 introduces a set of utilities to quantitatively measure and depict separations between classes in model scores [48]. These utilities are based on the Mahalanobis distance, a multivariate meter stick [16] that takes distributional properties of each class into account. Using these utilities, analysts may confidently make statistical arguments about distances between experimental classes in a dataset.

Finally, the possibility of the existence a new fundamental electronic interaction (the $n - \pi^*$ interaction) in proteins was explored in Chapter 11. Given the massive amounts of chemical and biological data available in depositories like the protein databank (PDB, [4]) and the biological magnetic resonance bank (BMRB, [43]), cheminformatic and bioinformatic data mining efforts such as the one performed in Chapter 11 are becoming increasingly possible. Efforts to curate the existing wealth of data into usable quantities, such as models of protein ^{1}H, ^{13}C and ^{15}N chemical shifts [37, 39, 26, 21, 27] and order parameters [2, 3] have proven useful in protein structure determination and refinement protocols. Combined with high-accuracy distance [44] and orientation [28] restraints, these databases will increasingly serve as sources of prior information in novel probabilistically driven structure determination efforts [34, 36].

While the raw amount of data is not quite as overwhelming in chemometrics as it may be in cheminformatic and bioinformatic studies, its complexity is just as great. As efforts to model protein structure and dynamics, cellular metabolic flux, and multi-component chemical mixtures continue, data handling methods must be advanced to ensure maximal "information handling" is achieved.

These methods must be provided in easy-to-use software packages that allow their users to focus on scientific inquiry, rather than trudging though manual pages to find the special syntax of a given function. Of course, doing so will require deep, multidisciplinary collaboration between groups that specialize in computer science, mathematics, statistics, chemistry and biology.

12.2 References

[1] W. Astle, M. de Iorio, S. Richardson, D. Stephens, and T. Ebbels. A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures. *Journal of the American Statistical Association*, 107(500):37–41, 2012.

[2] M. V. Berjanskii and D. S. Wishart. A Simple Method to Predict Protein Flexibility Using Secondary Chemical Shifts. *Journal of the American Chemical Society*, 127:14970–14971, 2005.

[3] M. V. Berjanskii and D. S. Wishart. Application of the random coil index to studying protein flexibility. *Journal of Biomolecular NMR*, 40:31–48, 2008.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[5] J. Boccard and D. N. Rutledge. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Analytica Chimica Acta*, 769:30–39, 2013.

[6] R. G. Brereton. A short history of chemometrics: A personal view. *Journal of Chemometrics*, 28(10):725–736, 2014.

[7] G. L. Bretthorst. Bayesian Analysis I. Parameter Estimation Using Quadrature NMR Models. *Journal of Magnetic Resonance*, 88:533–551, 1990.

[8] G. L. Bretthorst. Bayesian Analysis II. Signal Detection and Model Selection. *Journal of Magnetic Resonance*, 88:552–570, 1990.

[9] G. L. Bretthorst. Bayesian Analysis III. Applications to NMR Signal Detection, Model Selection and Parameter Estimation. *Journal of Magnetic Resonance*, 88:571–595, 1990.

[10] M. Bylesjo, R. Nilsson, V. Srivastava, A. Gronlund, A. I. Johansoon, S. Jansson, J. Karlsson, T. Moritz, G. Wingsle, and J. Trygg. Integrated Analysis of Transcript, Protein and Metabolite Data to Study Lignin Biosynthesis in Hybrid Aspen. *Journal of Proteome Research*, 8(1):199–210, 2009.

[11] R. A. Chylla, K. Hu, J. J. Ellinger, and J. L. Markley. Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: Application to quantitative metabolomics. *Analytical Chemistry*, 83(12):4871–4880, 2011.

[12] R. A. Chylla and J. L. Markley. Improved frequency resolution in multidimensional constant-time experiments by multidimensional Bayesian analysis. *Journal of Biomolecular NMR*, 3:515–533, 1993.

[13] R. A. Chylla and J. L. Markley. Theory and application of the maximum likelihood principle to NMR parameter estimation of multidimensional NMR data. *Journal of Biomolecular NMR*, 5(3):245–258, 1995.

[14] R. A. Chylla, B. F. Volkman, and J. L. Markley. Practical model fitting approaches to the direct extraction of NMR parameters simultaneously from all dimensions of multidimensional NMR spectra. *Journal of Biomolecular NMR*, 12(2):277–297, 1998.

[15] A. de Juan and R. Tauler. Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. *Critical Reviews in Analytical Chemistry*, 36(3-4):163–176, 2006.

[16] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[17] C. D. Eads, C. M. Furnish, I. Noda, K. D. Juhlin, D. A. Cooper, and S. W. Morrall. Molecular Factor Analysis Applied to Collections of NMR Spectra. *Analytical Chemistry*, 76(7):1982–1990, 2004.

[18] J. W. Eaton, D. Bateman, and S. Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.

[19] M. T. Eddy, D. Ruben, R. G. Griffin, and J. Herzfeld. Deterministic schedules for robust and reproducible non-uniform sampling in multidimensional NMR. *Journal of Magnetic Resonance*, 214:296–301, 2012.

[20] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre. A tutorial review: Metabolomics and partial least squares discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879:10–23, 2015.

[21] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart. SHIFTX2: Significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50:43–57, 2011.

[22] M. Hanafi and H. A. L. Kiers. Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Computational Statistics and Data Analysis*, 51:1491–1508, 2006.

[23] J. Hao, W. Astle, M. de Iorio, and T. M. D. Ebbels. BATMAN: An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15):2088–2090, 2012.

[24] K. Hu, J. J. Ellinger, R. a. Chylla, and J. L. Markley. Measurement of absolute concentrations of individual compounds in metabolite mixtures by gradient-selective time-zero ^{1}H–^{13}C HSQC with two concentration references and fast maximum likelihood reconstruction analysis. *Analytical Chemistry*, 83(24):9352–9360, 2011.

[25] S. G. Hyberts, K. Takeuchi, and G. Wagner. Poisson-Gap Sampling and Forward Maximum Entropy Reconstruction for Enhancing the Resolution and Sensitivity of Protein NMR Data. *Journal of the American Chemical Society*, 132(7):2145–2147, 2010.

[26] M. Iwadate, T. Asakura, and M. P. Williamson. C^{α} and C^{β} carbon-13 chemical shifts in proteins from an empirical database. *Journal of Biomolecular NMR*, 13:199–211, 1999.

[27] D.-W. Li, D. Meng, and R. Bruschweiler. Reliable resonance assignments of selected residues of proteins with known structure based on empirical NMR chemical shift prediction. *Journal of Magnetic Resonance*, 254:93–97, 2015.

[28] F. Li, J. Lee, A. Grishaev, J. Ying, and A. Bax. High Accuracy of Karplus Equations for Relating Three-bond J-couplings to Protein Backbone Torsion Angles. *ChemPhysChem*, 16(3):572–578, 2015.

[29] T. Lofstedt and J. Trygg. OnPLS – a novel multiblock method for the modeling of predictive and orthogonal variation. *Journal of Chemometrics*, 25:441–455, 2011.

[30] D. D. Marshall, S. Lei, B. Worley, Y. Huang, A. Garcia-Garcia, R. Franco, E. D. Dodds, and R. Powers. Combining DI-ESI-MS and NMR datasets for metabolic profiling. *Metabolomics*, 11(2):391–402, 2015.

[31] M. Mobli. Reducing seed-dependent variability of non-uniformly sampled multidimensional NMR data. *Journal of Magnetic Resonance*, 256:60–69, 2015.

[32] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *IEEE Transactions on Signal Processing*, 54(11):4133–4145, 2006.

[33] S. Moussaoui, C. Carteret, D. Brie, and A. Mohammad-Djafari. Bayesian analysis of spectral mixture data using Markov Chain Monte Carlo Methods. *Chemometrics and Intelligent Laboratory Systems*, 81(2):137–148, 2006.

[34] M. Nilges, A. Bernard, B. Bardiaux, T. Malliavin, M. Habeck, and W. Rieping. Accurate NMR Structures Through Minimization of an Extended Hybrid Energy. *Structure*, 16:1305–1312, 2009.

[35] M. F. Ochs, S. Stoyanova, F. Arias-Mendoza, and T. R. Brown. A New Method for Spectral Deconvolution Using a Bilinear Bayesian Approach. *Journal of Magnetic Resonance*, 137:161–167, 1999.

[36] S. Olsson, B. R. Vogeli, A. Cavalli, W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, and T. Hamelryck. Probabilistic Determination of Native State Ensembles of Proteins. *Journal of Chemical Theory and Computation*, 10:3484–3491, 2014.

[37] K. Osapay and D. A. Case. A New Analysis of Proton Chemical Shifts in Proteins. *Journal of the American Chemical Society*, 113:9436–9444, 1991.

[38] A. K. Smilde, J. A. Westerhuis, and S. de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337, 2003.

[39] S. Spera and A. Bax. Empirical correlation between protein backbone conformation and Calpha and Cbeta ^{13}C nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society*, 113(14):5490–5492, 1991.

[40] R. Stoyanova, J. K. Nicholson, J. C. Lindon, and T. R. Brown. Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets. *Analytical Chemistry*, 76(13):3666–3674, 2004.

[41] H. S. Tapp and E. K. Kemsley. Notes on the practical utility of OPLS. *Trends in Analytical Chemistry*, 28(11):1322–1327, 2009.

[42] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[43] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, W. R. Kent, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:402–408, 2008.

[44] B. R. Vogeli. The nuclear Overhauser effect from a quantitative perspective. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 78:1–46, 2014.

[45] J. A. Westerhuis and P. M. J. Coenegracht. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11(5):379–392, 1997.

[46] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[47] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[48] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, 2013.

[49] B. Worley and R. Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

[50] B. Worley and R. Powers. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[51] B. Worley and R. Powers. Simultaneous phase and scatter correction for NMR datasets. *Chemometrics and Intelligent Laboratory Systems*, 131:1–6, 2014.

[52] B. Worley and R. Powers. A Sequential Algorithm for Multiblock Orthogonal Projections to Latent Structures. *Chemometrics and Intelligent Laboratory Systems*, 2015.

[53] B. Worley and R. Powers. Deterministic Multidimensional Nonuniform Gap Sampling. *Journal of Magnetic Resonance*, 2015.

[54] B. Worley and R. Powers. Generalized Adaptive Intelligent Binning of Multiway Data. *Chemometrics and Intelligent Laboratory Systems*, 146:42–46, 2015.

[55] B. Worley and R. Powers. PCA as a predictor of OPLS-DA model reliability. *Analytica Chimica Acta*, 2015.

[56] B. Worley, N. J. Sisco, and R. Powers. Statistical Removal of Background Signals from High-throughput ^{1}H NMR Line-broadening Ligand-affinity Screens. *Journal of Biomolecular NMR*, 63(4):53–58, 2015.

[57] C. Zheng, S. C. Zhang, S. Ragg, D. Raftery, and O. Vitek. Identification and quantification of metabolites in ^{1}H NMR spectra by Bayesian model selection. *Bioinformatics*, 27(12):1637–1644, 2011.